5-2016

A Grounded Theory Exploration of the North Carolina Educator Evaluation System and its Effects on Teaching Practices and Teacher Leadership

Daniel A. Wydo
Gardner-Webb University

A Grounded Theory Exploration of the North Carolina Educator Evaluation System and its Effects on Teaching Practices and Teacher Leadership

By
Daniel A. Wydo

A Dissertation Submitted to the
Gardner-Webb School of Education
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Education

Gardner-Webb University
2016

Approval Page

This dissertation was submitted by Daniel A. Wydo under the direction of the persons listed below. It was submitted to the Gardner-Webb University School of Education and approved in partial fulfillment of the requirements for the degree of Doctor of Education at Gardner-Webb University.

_____ _____

Doug Eury, Ed.D. Date
Committee Chair

_____ _____

Bruce Boyles, Ed.D. Date
Committee Member

_____ _____

Jeff Hamilton, Ed.D. Date
Committee Member

_____ _____

Jeffrey Rogers, Ph.D. Date
Dean, Gayle Bolt Price School
of Graduate Studies

Abstract

A Grounded Theory Exploration of the North Carolina Educator Evaluation System and its Effects on Teaching Practices and Teacher Leadership. Wydo, Daniel A., 2016: Dissertation, Gardner-Webb University, Teacher Evaluation Systems/Value-Added/ Validity/Reliability/Unintended Consequences/Formative Evaluation/Summative Evaluation

This study examined the effects of the recently implemented North Carolina Educator Evaluator System (NCEES) on teaching practices and teacher leadership in a mostly rural county in the Piedmont region of North Carolina. NCEES is designed to improve teaching practices and teacher leadership through performance-based standards.

This mixed-methodology study began using grounded theory to form categories from qualitative data collected from piloted focus groups and interviews. Categories derived from the grounded theory analysis were refined in a secondary research method guided by a historical analysis of the processes related to teacher evaluation systems across many decades. The refined categories were then used to drive the primary research methods. A questionnaire was developed based on the refined categories; checked for construct, content, and item validity and reliability; and distributed as part of a survey process in the county under study. The questionnaire was designed using a Likert scale to measure teachers' perceptions of the effects of NCEES on teaching practices and teacher leadership. The questionnaire also allowed teachers' written responses on unstructured questions for which grounded theory was used to analyze. In order to triangulate, aggregate teacher ratings from NCEES were examined quantitatively to detect the effects of NCEES on teaching practices and teacher leadership.

Ninety teachers and six administrators participated in the primary research and supplied substantive qualitative data. Quantitative data for the primary research were extracted online from the North Carolina Department of Public Instruction.

The study found conflicting evidence of an overall effect of NCEES on teaching practices due to the inability of principals and teachers to identify specific improvements, conflicts in the quantitative and qualitative analysis of NCEES Teacher Survey, and conflicts within the distribution of teacher rankings across years and compared to three observed sources from the literature review. However, it was determined that teacher leadership had improved due to NCEES based on specific responses from principals and teachers across the Principal Interview process and NCEES Teacher Survey. Evidence was uncovered that indicated teachers primarily improved their teacher leadership roles within professional learning communities and school improvement activities.

Table of Contents

Figures

Chapter 1: Introduction

Statement of the Problem

People remember great teachers and for good reasons. Raymond (2007) reflected upon the teaching practices and teacher leadership roles of his three favorite teachers:

> We all remember our favorite teachers and we probably all agree they were good. However–an interesting thought–were they *great* teachers? And if so, what's the magic that made those good teachers great?
>
> Did they make us laugh? Did they work us hard? Did their enthusiasm for the courses they taught wash over us–making us enthusiastic, too?
>
> The classrooms of my three favorite teachers were stimulating, lively, challenging, exciting, and now, over a half-century later, I still remember where I sat in their classrooms. There is no doubt about it–in my mind they were great teachers. (p. 6)

Raymond (2007) revered his favorite teachers and informally identified why they were so memorable. However, there have been a number of formal attempts to understand why some teaching practices and forms of teacher leadership were memorably successful. Researchers in education have worked to discover and replicate teaching practices and forms of teacher leadership that provided the "stimulating," "lively," "challenging," and "exciting" classroom environment that Raymond memorialized (Measures of Effective Teaching [MET], 2010, 2012, 2013).

Many studies, such as *A Nation at Risk: The Imperative for Educational Reform*, have supplied federal and state policymakers with data that have driven reform movements in education (Gardner, 1983). Reform legislation, such as the Race to the Top (RttT) initiative enacted in 2009 (U.S. Department of Education [USDOE], 2014), Goals

2000 enacted in 1994 (USDOE, n.d.a), and the Elementary and Secondary Education Act (ESEA) originally enacted in 1965 (Federal Education Policy History, 2011) and reauthorized and enacted in 2002 as the No Child Left Behind (NCLB) Act (USDOE, 2010b), have sought to increase student achievement by improving teaching practices and teacher leadership. An increasingly popular topic in education has been investigating successful teachers and how they teach and lead in supporting student learning (Auguste, Kihn, & Miller, 2010; Gordon, Kane, & Staiger, 2006). The federal and state governments have become more actively involved in studying successful teachers and concurrently teaching practices and leadership abilities (Government Accountability Office [GAO], 2009).

GAO (2009) found that over a third of the programs that the USDOE financed hinged on "efforts to improve teacher quality" (p. 11). GAO (2009) also found that between February 2008 and July 2009, nine offices of the federal government allocated billions of federal dollars to improve teaching practices and teacher leadership through statutorily authorized programs.

The GAO (2009) study emphasized the need for federal and state governments to design and sustain coordination among program offices in order to enhance the limited resources that state education officials reported available to them. GAO (2009) surveyed state departments and found that the limited resources related mostly to the lack of funds necessary to build an infrastructure that would track the quality of teachers throughout systems.

The trend at the federal level concentrated on improving the quality of teachers as a previously existing $65 million dollar federal project used to track teacher quality received another $250 million dollars in the American Recovery and Reinvestment Act of

2009 (ARRA) (GAO, 2009). The purpose of the additional $250 million dollars from the ARRA required states to report on the progress they were making toward the design of statewide data systems (GAO, 2009). GAO (2009) described future data systems as mechanisms that would allow the matching of individual students to specific teachers in order to measure the quality of teaching practices and teacher leadership.

In 2006, Congress enacted the Teacher Incentive Fund (TIF) with $600 million and continued TIF with funds of $440 million in 2009 under the ARRA of 2009 (Glazerman, Chiang, Wellington, Constantine, & Player, 2011). The purpose of the TIF was to identify and reward teachers based on effective teaching practices and teacher leadership (Glazerman et al., 2011).

At the federal and state levels, legislation has been in response to outcomes on standardized tests such as the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and the Program for International Student Assessment (PISA) (Schneider, 2009). Both the TIMSS and PISA have shown the United States is lagging behind other countries in mathematics, science, and reading. The U.S. ranked 14th in reading, 17th in science, and 25th in math on the 2009 PISA (Fleischman, Hopstock, Pelczar, Shelley, & Xie, 2010). The U.S. ranked 11th on the fourth-grade 2007 TIMSS mathematics examination and ninth on the eighth-grade 2007 TIMSS mathematics examination (Gonzales, Williams, Jocelyn, Roey, Kastberg, & Brenwald, 2009).

The NAEP standardized test has served as an internal source for U.S. student achievement and has allowed for a comparison from state to state. Although there have been improvements in NAEP outcomes for mathematics over the last 20 years, reading levels have remained flat (Paulson, 2011). Paulson (2011) found that few students have

reached the "proficient" level set by NAEP.

According to the National Center for Education Statistics (NCES), in 2011, 34% of both fourth and eighth graders scored at or above "proficient" on the NAEP reading assessment; 40% and 35% of fourth and eighth graders scored at or above "proficient' on the NAEP mathematics assessment, respectively; and 32% of eighth graders scored at or above "proficient" on the NAEP science assessment (fourth-grade science NAEP assessment was not administered in 2011) (NCES, 2011a, 2011b, 2011c). NCES (2001a, 2011b, 2011c) data reinforced Paulson's (2011) insight that few students are reaching the NAEP's cut score for proficiency.

There have been a number of legislative acts to counteract the substandard standardized test scores, including the NCLB Act of 2002 and the RttT initiative of 2009 (USDOE, 2010b, 2014). Much of this legislation transformed the educational landscape in an attempt to discover and replicate how successful schools and teachers advance learning in terms of teaching practices and teacher leadership (Lohman, 2010). The RttT initiative of 2009 was an important piece of recent legislation that affected teaching practices and teacher leadership (USDOE, 2009, 2014).

In 2009, U.S. Secretary Arne Duncan and President Barack Obama introduced the 4.35 billion dollar RttT initiative. Rather than divvying out reform money to states, the federal government required that states compete over the funding (USDOE, 2009). The states applied for the money over the course of three rounds spanning from January 19, 2010 to December 23, 2011 (USDOE, 2011b).

Secretary Duncan made it known that the money made available through the ARRA was the largest discretionary amount of money ever dedicated to school reform when he said,

> if you take all the discretionary money for reform that every one of my predecessors had and add it all together for the last 29 years in a row; it is still a much smaller pot of money for reform than the Race to the Top fund that we're announcing today. (Duncan & Obama, 2009)

While still in its infancy, RttT dictated that states concentrate their reform efforts in the following four areas given by the USDOE (2009):

- Adopting standards and assessments that prepare students to succeed in college and the workplace and to compete in the global economy;
- Building data systems that measure student growth and success, and inform teachers and principals about how they can improve instruction;
- Recruiting, developing, rewarding, and retaining effective teachers and principals, especially where they are needed most; and
- Turning around our lowest-achieving schools. (p. 2)

RttT grant funds aimed to improve teaching practices and teacher leadership in a variety of ways based on competitive state applications. Redesigning teacher evaluation systems at the state level was one way for states to show the federal RttT application assessors that they were committed to improving teaching practices and teacher leadership (USDOE, 2009).

One aspect of the RttT application process dominated the four areas of the RttT initiative–the Great Teachers and Leaders area. As the National Council on Teacher Quality (NCTQ, 2009) recognized,

> The Great Teachers and Leaders area of the application is a clear make or break for states. It is the most important single assurance area in terms of points assigned. It alone can earn a state 138 points out of the 500 total points for the

application, more than a quarter of the entire RttT point allocation. (p. 1)

Hanover Research (2011) also stated,

> Though RttT targeted various levels of reform, the most significant emphasis was placed on improving teacher effectiveness, specifically by bolstering evaluation procedures. This encompassed two main dimensions: first, establishing a means of measuring student growth (ideally through the implementation of data systems), and second, incorporating this data into teacher evaluations and subsequent human capital decisions, such as promotion, retention, and tenure. (p. 2)

In the 2009 speech to the National Education Association (NEA), Secretary Duncan labeled the nation's teacher evaluation systems as flawed resulting in students who lose out on achievement opportunities (Duncan, 2009). An important component of ensuring a high level of student achievement has been to examine teaching practices and teacher leadership roles in producing learning environments within the classroom (Schmoker, 2006).

The federal government included this approach as an important facet of the RttT funding competition (USDOE, n.d.b). Quality teacher evaluation systems could serve as catalysts to improve teaching practices and teacher leadership, uncover weaknesses for future professional development offerings, and identify and retain highly effective teachers to serve as mentors to inexperienced or struggling teachers–all of which can contribute to higher student achievement (Danielson, 2010/2011).

Regardless of the quantity or areas of reform, the importance of one variable has reoccurred–the quality of a teacher. Wenglinsky (2000) pointed out,

> As we have considered specific reforms as diverse as early childhood education,

> smaller class sizes, parental involvement, and learning technology, one point became abundantly clear. Unless a child is taught by competent teachers, the impact of other education reforms will be diminished. Simply put, students learn more from "good" teachers than from "bad" teachers under any set of circumstances. (p. 5)

It has been shown that the competency of the teacher has been the first and foremost factor that influences student learning within schools, specifically whether they implement quality teaching practices and effective leadership roles. Wright, Horn, and Sanders (1997) showed the effectiveness of teachers mattered more than other school-site variables such as the heterogeneity of students in the classrooms, class size, how students were assigned classes, or the school district in which students attended.

If highly effective teachers could be identified and isolated through capable teacher evaluation systems, there would be much to learn from them, and it would be possible to answer questions about how effective teachers operate as the Bill and Melinda Gates Foundation (2010) has investigated through the MET project. As the Bill and Melinda Gates Foundation (2010) pronounced,

> Evidence shows clearly what most people know intuitively: teachers matter more to student learning than anything else inside a school. Yet the design of our school systems fails to value and support effective teaching, improve overall teacher effectiveness, or recognize and reward those who take on and excel at the most difficult teaching assignments. Teachers often bear the brunt of the failure to recognize effective teaching. We do little to study and share the best teachers' practices or recognize them for their contributions. (p. 9)

The Bill and Melinda Gates Foundation (2010) also stated,

> There is no widely agreed upon measure for teacher effectiveness that exists today, and that is precisely why we are undertaking this work. The results of this project and what we learn will help districts across the country identify effective teaching in order to improve student achievement and help teachers ensure excellence in their profession. (p. 10)

The problem under investigation in this study surrounded the effects of teacher evaluation systems on teaching practices and teacher leadership. Specifically, this study was designed to determine how the North Carolina Educator Evaluation System (NCEES) has affected teaching practices and teacher leadership in six local high schools in the southwestern piedmont region of North Carolina. NCEES was formally implemented in the 2011-2012 school year.

Once the envy of the world, the U.S. educational system has struggled since the 1970s to produce quality graduates (Hershberg, 2005), while teacher evaluation systems have neglected to identify failing teachers (Weisberg et al., 2009). The goal in designing more robust teacher evaluation systems has been to place an emphasis on quality teachers who bring the best teaching practices and teacher leadership abilities to classrooms across the U.S. (Danielson, 2010/2011).

There has been evidence uncovered that the problems with public schools in the U.S., to some extent, stem from ineffective, perfunctory teacher evaluation systems (Danielson, 2010/2011). Historically, there has been research indicating that teachers have not chosen quality classroom teaching practices and have not displayed the type of professional leadership necessary to improve student achievement (Goodlad & Klein, 1970). During the late 1980s and early 1990s, Frase and Streshly (1994) analyzed the quality of teaching practices and teacher leadership roles of teachers and provided

evidence that there was a lack of quality teaching practices and teacher leadership. There continues to be evidence that teachers do not choose the best classroom practices and forms of leadership known to increase student learning (MET, 2012; Schmoker, 2006).

Studies by Goodlad and Klein (1970), Frase and Streshly (1994), Schmoker (2006), and MET (2012) were crucial in exposing a lack of quality teaching practices and teacher leadership. All four studies involved large representative samples (totaling thousands of classroom observations of random teachers) and were from different time periods.

Goodlad and Klein (1970) collected data retrieved by multiple researchers who collected data from "158 classrooms of 67 schools in 26 school districts" (p. 33). The study involved 13 states with a geographical spread (Goodlad & Klein, 1970). Goodlad and Klein concluded that "general or specific classroom goals were not identifiable to observers. Instruction was general in character and not specifically directed to diagnosed needs, progress, and problems of individual children. Teachers shot with a shotgun, not a rifle" (p. 98).

Frase and Streshly (1994) analyzed data collected by trained audit teams in six school districts spanning from 1987 to 1991, which involved teachers of over 195,000 students. Frase and Streshly cited results from auditors that "revealed very poor instructional practices" (p. 49) including the following list of "predominant examples" (p. 49) of teacher behavior:

1. Exclusive use of drill and practice exercises.
2. Students' copying exercises *and* directions from books and workbooks.
3. Teachers not engaged with students (e.g., teachers with heads on desks, grading papers, coloring posters, and doing other seat work, while students

completed rote exercises).

4. Students asleep or not being attentive in class.
5. Absence of lesson plans.
6. Use of low quality and non-stimulating worksheets.
7. Students copying words from the chalkboard and writing down their definitions from a dictionary.
8. Teachers asking questions printed in text and a very limited number of students raising their hands to answer.
9. Classes unattended by teachers. (pp. 49-50)

Schmoker (2006) relied on a study carried out by Learning 24/7 in 2005, along with his own insights, and took notes of the following shortcomings as evidenced from classroom observations (percentages illustrated frequency of behaviors that were observed):

- Classrooms in which there was evidence of a clear learning objective: 4%.
- Classrooms in which high-yield strategies were being used: 0.2%.
- Classrooms in which there was evidence of higher-order thinking: 3%.
- Classrooms in which students were either writing or using rubrics: 0%.
- Classrooms in which fewer than one-half of students were paying attention: 85%.
- Classrooms in which students were using worksheets (a bad sign): 52%.
- Classrooms in which non-instructional activities were occurring: 35%. (p. 18)

The MET (2012) project collected data from 1,333 teachers including videos of more than 7,491 lessons. The videos using TeachScape technology produced 22,000

observation ratings for teachers and involved more than 44,500 students. Based on multiple classroom observation instruments (CLASS, FFT, PLATO, MQI, and UTOP), MET (2012) concluded that

> Many of the lowest scores for UTOP competences related to skills associated with engaging students in rigorous instruction. Only about 35 percent of lessons scored a three or above on *intellectual engagement with key* ideas. The same was true for about 30 percent of lessons on questioning strategies and only about 15 percent on *investigation/problem-based* approach.
>
> These classroom observation instruments, though they differ in emphasis, portray a remarkably consistent image of classroom practice. The MET project teachers tended to do fairly well at behavior management and time management. However, scores were lower in areas such as problem solving (CLASS), effective discussion (FFT), intellectual challenge (PLATO), richness (MQI), and investigation (UTOP). The task of developing conceptual understanding is complex, and these results suggest that as the teaching tasks grew in complexity, it grew rare. (p. 25)

Goodlad and Klein (1970), Frase and Streshly (1994), Schmoker (2006), and MET (2012) provided evidence that there has been room for innovation and improvement in teaching practices and teacher leadership roles. These four studies involved multiple observations of teaching practices and teacher leadership across a wide geographical span within the U.S. school systems and from different time periods.

In analyzing the lack of quality in teaching practices and teacher leadership, authors often measure the output of teaching practices and teacher leadership in terms of the quality of the product–the U.S. high school graduate. Many studies have provided

evidence that high school graduates have not been ready for college or to enter the workplace. These studies included *The Condition of College and Career Readiness* carried out by the American College Testing (ACT, 2011); *Remediation–The Bridge to Nowhere* carried out by Complete College America (CCA, 2012); the *National Curriculum Survey* carried out by (ACT, 2009); *Are They Really Ready to Work* carried out by The Conference Board, Partnership for 21st Skills, Corporate Voices for Working Families & the Society for Human Resource Management (The College Board et al., 2006). The indicators of a lack of student achievement among U.S. students included the following:

1. The ACT indicator has provided evidence that high school students have not been ready for success in college (ACT, 2011).
2. There has been a high incidence rate of remediation for college freshman (CCA, 2012; USDOE, 2010a).
3. College professors have had a high level of agreement that high school students are not academically ready for college (ACT, 2009).
4. Employers in the workplace have had a high level of agreement that high school students are not ready to enter socially or academically (The Conference Board et al., 2006; USDOE, 2010a).

College graduates have earned higher lifetime incomes than high school graduates: Based on data from 1999, the average lifetime earnings for individuals with a high school diploma was $1.2 million versus $1.5 million for individuals with some college experience, $1.6 million for individuals with an associate's degree, $2.1 million for individuals with a bachelor's degree, $2.5 million for individuals with a master's degree, and $3.4 million for individuals with a doctorate degree (Day & Newburger,

2002). Teachers have purposed to incrementally prepare students ensuring high school graduates were college or career ready. The ACT has been an important college and career readiness indicator as stated by *The Condition of College and Career Readiness* (ACT, 2011): "ACT's College Readiness Standards are precise descriptions of the essential skills and knowledge that students need to become ready for college and career, beginning in grade 8 and continuing through grade 12" (p. 25).

According to ACT (2011), "25% of high school graduates met the college readiness requirements set in English, math, reading, and science," while "28% met none of the college readiness requirements" (p. 4). In response to the release of the ACT scores, Secretary Duncan (2011a) responded that "These ACT results are another sign that states need to raise their academic standards and commit to education reforms that accelerate student achievement" (para. 3).

Another crucial measure of student outcomes has been to track remediation rates of incoming freshman at the college and university levels. Although more high school graduates are enrolling and attending college than ever before (63% in 2009 compared to 54% in 1992) (National Center for Higher Education Management Systems, 2012), many freshmen require remediation (52% in 2-year colleges; 20% in 4-year colleges) (CCA, 2012).

There has been evidence that high school teachers differ in their perceptions, compared to college professors, regarding the college readiness of high school graduates. While high school teachers perceive they have been graduating high school students ready for college, college and university professors have disagreed (ACT, 2009). In the ACT (2009) *National Curriculum Survey*, 91% high school teachers reported that their students were prepared for college-level work, whereas 26% of college professors

reported their students arrive prepared.

Finally, according to The College Board et al. (2006), high school graduates have been "ill-prepared for the demands of today's (and tomorrow's) workplace" (p. 9). The study involved 400 employers who were asked about high school graduates who entered their respective fields; and the employers rated the new entrees as "deficient," "adequate," "excellent," or "NA" (The Conference Board et al., 2006). High school graduates were rated as "deficient" in basic knowledge of English, math, and reading; "deficient" in writing and critical thinking/problem solving; "deficient" in professionalism/work ethic; and "adequate" in technology applications and teamwork/collaboration (The Conference Board et al., 2006, p. 11).

The evidence that there was a lack of quality in teaching practices and teacher leadership from Goodlad and Klein (1970), Frase and Streshly (1994), Schmoker (2006), and MET (2012) paralleled high school graduates who were not ready to enter college per ACT (2011), who had high remediation rates as college freshmen per CCA (2012), and were overall "deficient" when entering the workplace per The Conference Board et al. (2006). However, how did the lack of teaching quality and teacher leadership and poor quality of products from the U.S. public school system compare to how teachers have been rated within teacher evaluation systems?

The evidence of the declining quality in teaching practices and teacher leadership, alongside the decline in the aptitude of U.S. high school graduates, did *not* correlate with teacher ratings from teacher evaluation systems. While teaching practices and teacher leadership roles and the quality of U.S. high school graduates have lacked in quality, teachers widely received high ratings within teacher evaluation systems across the U.S. (Frase & Streshly, 1994; Weisberg et al., 2009). Authors have noted the contradictory

evidence of high-performing teachers and low student outcomes (Gonzalez, 2012; Murphy, 2011; Weisberg et al., 2009). In most cases, 99% of teachers were found to be performing at "satisfactory" levels (Frase & Streshly, 1994; Weisberg et al., 2009).

The New Teacher Project (TNTP) showed that among 12 school districts and four states, 99% of teachers were rated as "satisfactory" when teachers were rated in a binary fashion ("satisfactory" or "unsatisfactory") (Weisberg et al., 2009, p. 6). In school systems where there were more than two ratings, 94% of teachers were rated with one of the highest categorical ratings, and less than 1% of teachers were rated as "unsatisfactory" (Weisberg et al., 2009). Furthermore, of approximately 15,000 teachers and 1,300 administrators involved in the study, 57% of teachers and 81% of administrators said there was a poorly performing, tenured teacher in their school; and 43% of teachers said they knew a poorly performing, tenured teacher who should be dismissed (Weisberg et al., 2009).

Examples of low student achievement and high teacher evaluation ratings abound. In Pennsylvania, teachers and administrators were rated as satisfactory, yet student achievement was lacking (Murphy, 2011). Pennsylvania State Education Secretary of Education Ronald Tomalis stated,

> It is very difficult for me to rationalize how our state can have virtually 100 percent of educators evaluated as satisfactory when, based on the statewide assessment, one-in-four students are scoring below proficient in reading, and one-in-three are scoring below proficient in math.
>
> What's more disturbing, based on the National Assessment of Educational Progress (NAEP), more than half of our fourth- and eighth-grade students are scoring below proficient in math and reading. I believe these results are a clear

> indication that our current evaluation system is in major need of change. (Murphy, 2011, para. 2)

Duncan (2010) added, "Everyone agrees that our teacher evaluation system is broken. In many districts, 99% of teachers are rated satisfactory and most evaluations ignore the most important measure of a teacher's success–which is how much their students have learned" (para. 1). Cody, McFarland, Moore, and Preston (2010) noted, in conjunction with Eckert and Dabrowski (2010),

> At the same time, most education officials acknowledge that the current teacher evaluation and tenure process does not place enough emphasis on actual student achievement; far less than 90 percent of American students are receiving a quality education even though over 90 percent of teachers in the nation receive "satisfactory ratings" on teacher evaluations. (p. 8)

There have been years of reform efforts aimed at teacher quality in the state of Indiana (Ladner, Lips, & Daniels, 2012). According to Ladner et al. (2012), those reform efforts over the years had only wrought partial success until Indiana passed landmark legislation emphasizing "teacher quality, administrative flexibility, school accountability, and parent and student choice" (p. vii). Mitch Daniels, the Governor of Indiana, as part of the American Legislative Council (ALEC) report, remarked,

> Prior to this session, 99 percent of Indiana's teachers were annually rated "Effective." If that rating were actually true, 99 percent–not just one-third–of our students would be passing national tests. From this point on, because of the diligence and fortitude of our reform-minded legislators, teachers will be promoted and retained based on performance rather than seniority. (Ladner et al., 2012, p. vii)

From the Frase and Streshly (1994) study, of the six school districts and 195,000 students involved, very few teachers were rated "below standard." In one of the largest districts involved in the study, there were 40,000 students under teachers being observed; and Frase and Streshly noted that there were no teachers found "below standard" when observed by an administrator, regardless of external auditors finding "very poor instructional practices" (p. 49). Although few teachers were rated at "below standard" by administrators, external auditors found teachers exclusively relying on drill and practice exercises, students asleep or not being attentive in class, that teachers did not have lesson plans, and that there were teachers not in their classrooms while students were (Frase & Streshly, 1994).

Teacher evaluation data and student learning outcomes have not aligned historically. If teaching practices and teacher leadership roles were highly rated by administrators within teacher evaluation systems (Weisberg et al., 2009), then why have high school graduates failed to meet the mark of college readiness (ACT, 2011), failed to meet employment requirements (CCA, 2012), and lacked in comparison to graduates from other countries (Hershberg, 2005)? Also, why did the external auditors in the Frase and Streshly (1994) study widely find teachers not using the best teaching practices and forms of leadership known to increase student learning while the administrators in the six school districts rarely rated teachers "below standard"?

Background of the Problem

In order to cultivate quality teaching practices and teachers as leaders, North Carolina designed a new teacher evaluation system, NCEES, and piloted it during the 2009-2010 school year. NCEES was officially enacted statewide during the 2010-2011 school year. NCEES was based on a rubric for qualitatively driven administrator ratings,

developed by the Mid-Continent Research for Education and Learning (McREL), the North Carolina State Board of Education (NCSBOE), and the North Carolina Department of Instruction (NCDPI). NCEES has also relied on the Statistical Analysis System Institute (SAS) for the quantitatively based value-added modeling (VAM). VAM has been used within NCEES to provide student growth outcomes mandated by the RttT initiative.

After using a 2-decade-old teacher evaluation instrument, the Teacher Performance Appraisal Instrument (TPAI) and the TPAI-Revised (TPAI-R), and formulating a new vision for learning, the NCSBOE required that

> A new vision of school leadership and a new set of skills that teachers must use daily in order to help their students learn 21st century content and master skills they will need when they graduate from high school and enroll in higher education or enter the workforce or the military. (McREL, NCDPI, & NCSBOE, 2012, p. 4)

Based on this new vision, the former evaluation instrument was replaced in favor of a new system to "promote and support effective leadership, quality teaching, and student learning" (McREL et al., 2012, p. 4). The NCEES was to guide the process of assessing teaching practices and teacher leadership and was designed to give teachers the opportunity to grow professionally (McREL et al., 2012). There was a focus to develop teaching practices and teacher leadership built into the new NCEES (McREL et al., 2012), whereas the concentration previously was centered more on judging competencies (Milner, 1991).

Because NCEES recently has been implemented, there has been little research available in terms of its formative effects to help teachers develop quality teaching

practices and teacher leadership. The data from the TPAI or TPAI-R were not found in database form but followed teachers in personnel files.

In contrast, data from NCEES were housed statewide in a database that allowed state and county personnel to analyze teacher evaluation data and make professional development decisions based on deficiencies in the teacher evaluation data (McREL et al., 2012). These banked data have served advantageously in researching this topic to gauge the effects of NCEES on teaching practices and teacher leadership.

In considering the problems associated with a lack of quality in teaching practices and teacher leadership, Schmoker (2006) promoted a simple first step to evaluating teachers and thereby increasing student achievement–touring classrooms and observing student learning behaviors. Schmoker (2006) detailed that "a single one-hour tour was the catalyst for actions that resulted in dramatic, sustained gains in student achievement at elementary schools" (p. 16) within a California school district. This approach directly contradicted traditional approaches to evaluating and observing teachers where teachers experienced limited collaboration and relied on their own intuition and training (Huffman & Hipp, 2003).

Successful teachers have not historically collaborated to share teaching practices and forms of leadership that could transform struggling teachers, thereby increasing student achievement (Huffman & Hipp, 2003). Teachers have been "encapsulated" (Sarason, 1996, p. 329) in their own classrooms, disconnected from administrative purview, from peer review of fellow teachers, and from viable solutions that could improve their teaching practices and leadership roles.

Sarason's (1996) "encapsulation" had a cause and an effect as Schmoker (2006) suggested–teachers were not observed and evaluated consistently by administrators or

peers (Eaker, DuFour, & DuFour, 2002). As a result, successful teaching practices and teacher leadership went unnoticed and unused as a means to professionally develop faculty (Elmore, 2000). Eaker et al. (2002) noted, "Traditional schools are characterized, to a great degree, by teacher isolation. It has been said that the traditional school often functions as a collection of independent contractors united by a common parking lot" (p. 10).

Schmoker (2006) made a pronounced call for a closer inspection of teaching practices and forms of teacher leadership. According to Schmoker (2006), by concentrating on the observation and evaluation of teaching and learning within the classroom environment, observers could help improve teaching practices and leadership roles among teachers and achieve unprecedented improvements in student learning as a result.

Schmoker (2006) described a situation in which if one were to visit classrooms randomly and unannounced, he/she would encounter "the alarming gap between reality and our assumption about the general quality of classroom practices" (p. 15). Upon random, unannounced visits, observers would more than likely find vast amounts of busywork without reference to intended curriculum, the inability of teachers or students to articulate what they were to be teaching or learning that day, irrelevant worksheets and activities, and unengaged students (Schmoker, 2006). Schmoker (2006) added the following remarks in conjunction with making random, unannounced observations of teaching practices and teacher leadership roles in the classroom environment:

> And in defiance of what every educator has learned, there is a glaring absence of the most basic elements of an effective lesson: an essential, clearly defined learning objective followed by careful modeling or a clear sequence of steps,

> punctuated by efforts during the lesson to see how well students are paying attention or learning the material. In most classrooms, half or more of the students are clearly not engaged or paying attention. (p. 16)

Schmoker (2006), connecting with an apparent lack of quality in teaching practices and teacher leadership, also contended, "In general, there is very little oversight of instruction that affects its quality. Administrators do not have any common, formal mechanisms to accurately gauge the content teachers are actually teaching or how effectively they are teaching it" (p. 18). Such a lack in effective teacher practices and teacher leadership and such an inability of observers to evaluate and observe such processes begs the question, "What can be done to improve the situation?"

Elmore (2000) advised administrators directly manage instructional practices rather than concentrating on the traditional "structures and processes around instruction" (p. 6), such as structuring a bell schedule, caring for school grounds, or planning bus routes. Elmore viewed the traditional roles of principals as important, but also stressed an overall devotion to improving teaching practices and teacher leadership roles occurring in the classroom.

An ongoing concern in public schools has been the public being "buffered" from the quality of teaching practices and leadership within the classroom and schools by administrators who protect mediocre or subpar processes (Elmore, 2000, p. 2). Elmore (2000) argued that school administrators "buffer" or "protect" the classroom environment by giving the public "the impression that they are managing it" (p. 2), rather than dissecting the classroom environment and relaying the quality of instruction.

Schmoker (2006), Sarason (1996), and Elmore (2000) collectively described potential causes and solutions that educational researchers have identified with teacher

evaluation systems which lead to compromised teaching practices and a lack of teacher leadership. The "encapsulation" of solitary teachers identified by Sarason (1996), Schmoker's (2006) discovery of the lack of consistent and meaningful classroom observations and evaluation by administrators, and the lack of transparency as a "buffer" described by Elmore (2000), have been issues that potentially propagated poor to mediocre teaching practices and a lack of teacher leadership roles. Ineffective teacher evaluation systems have historically harbored such outcomes (Peterson, 2000; Weisberg et al., 2009). In the end, it has been the students who have paid the price as their achievement has been sacrificed due to ineffective evaluation systems.

Although Sarason (1996), Schmoker (2006), and Elmore (2000) advised increasing classroom observations and evaluation and openly transmit the qualitatively driven results for the sake of improvement, there has been a recently popularized quantitative form of accountability for teachers known as VAM (Bianchi, 2003; Kupermintz, 2003; Sanders & Horn, 1995; Sanders & Rivers, 1996; Wright, White, Sanders, & Rivers, 2010). VAM has measured the effectiveness of teachers based on the value of learning the teacher added to the academic growth of a student between two or more points in time (Bianchi, 2003; Hershberg, 2005; Kupermintz, 2003). The process of measuring teacher quality and effectiveness through student growth is not new, dating back to Hanushek (1972) and Murnane (1975); however, there has been a recent upshift in its use (Braun, Chudowsky, & Koenig; 2010; McCaffrey & Lockwood, 2008).

William L. Sanders used data from the Tennessee Comprehensive Assessment Program (TCAP) in the 1990s to statistically identify factors affecting student learning while controlling for a myriad of extraneous variables. In doing so, Sanders reliably estimated the effects of a teacher on the learning gains of students on standardized tests

over the course of time (Sanders & Horn, 1995; Wright et al., 1997). The teacher effect that Sanders isolated has been used to help evaluate teachers in terms of how much their students grew based on the difference between projected and observed test scores (Sanders & Horn, 1998; Sanders & Rivers, 1996).

Chetty, Friedman, and Rockoff (2011) investigated the effects of highly effective teachers and described their long-term benefits for students later in life based on the effectiveness of the teachers calculated by VAM. One standard deviation higher in VAM outcomes for teachers translated into higher average annual earnings for their students and a higher likelihood that the students would graduate from high school and attend college (Chetty at al., 2011). Chetty et al. also used VAM outcomes to show a positive correlation between highly effective teachers and students who had a lower likelihood of teenage pregnancy, were more likely to attend a higher quality college, and would later live in a higher quality neighborhood based on zip code analysis.

When used within teacher evaluation systems, VAM has changed the way teachers were evaluated by interjecting a quantitative, statistical methodology free from the bias of the observer effect and halo effect that plagued qualitative, classroom observation components of teacher evaluation systems (Goe, 2008). VAM has considered the amount of student academic progress over multiple school years rather than a measurement at a point in time that identifies student proficiency status such as "basic," "proficient," or "advanced" (Braun et al., 2010; Cody et al., 2010). Although administrators and classroom observers can be trained extensively on how to qualitatively recognize the best teaching practices and forms of leadership for teachers within classrooms, Goe (2008) recognized, "observer bias can be minimized but not eliminated; with value-added models, there is no observer–only scores" (p. 5).

The implementation of school-wide growth models have been utilized by some states within regulations appropriated by the NCLB Act and through the Growth Model Pilot Program (GMPP) from the 2006-2007 school year and beyond. The GMPP concentrated on student growth at the school level and was used for calculating and reporting Adequate Yearly Progress (AYP) (Braun et al., 2010). The RttT initiative concentrated on student growth at the level of teachers rather than the level of schools, and VAM has played a key role in this approach (Braun et al., 2010; USDOE, 2009).

In the past, teacher effectiveness was based primarily on subjective judgments about pedagogical ability and other outward characteristics such as certification status, college major, number of advanced degrees, and experience (MET, 2012). However, an observable shift occurred as federal reform legislation took root, specifically the NCLB Act of 2002 and the RttT initiative of 2009 (Lohman, 2010). Student growth began to dominate individualized accountability for teachers (Braun et al., 2010; Lohman, 2010).

The NCLB Act concentrated on reforming schools based on proficiency data from year to year (Braun et al., 2010; Lohman, 2010). This involved measuring the proficiency of groups (cohorts) from year to year and relied on formulas within AYP rules as a measuring stick to determine year-to-year growth among cohorts (Braun et al., 2010; Lohman, 2010). Another goal of the NCLB Act was to close the achievement gap between White and minority students (Braun et al., 2010; Lohman, 2010). The RttT of 2009 initiative was centered on teacher-level reform that identified successful teachers as those whose teaching practices and teacher leadership increased student growth, regardless of the students' prior proficiency status and socioeconomic factors (Hull, 2011; Lohman, 2010).

The linkage of teacher and student data became an integral part of teacher

evaluation systems and was implemented by states to compete in the RttT initiative (Hull, 2011; Lohman, 2010). The RttT initiative mandated longitudinal systems, which linked students with teachers, tracked the students' progress, and were to be used to measure the effectiveness of teachers and schools based on the students' growth on standardized tests scores (USDOE, 2009).

Forty states and the District of Columbia applied for funding in the RttT initiative; the winners of round one were Delaware and Tennessee as announced in March 2010; and the winners of round two were California (dropped due to incomplete application), the District of Columbia, Florida, Georgia, Hawaii, Maryland, Massachusetts, Minnesota, New York, North Carolina, Ohio, Rhode Island, and Washington as announced in August 2010 (USDOE, 2011b). Another seven states won round three of the RttT initiative: Arizona, Colorado, Illinois, Kentucky, Louisiana, New Jersey, and Pennsylvania as announced in December 2011 (USDOE, 2011b). Boser (2012) found that as a result of the RttT initiative, "States have also used their RttT dollars to create new teacher- and principal-evaluation systems that include student learning as a component, and our research found that all of the states have either piloted or implemented new teacher-evaluation systems" (p. 3).

Although about one-half of the states that competed won RttT money (through all three rounds), the states that did not win RttT money also made sweeping changes in order to make their applications more competitive (Boser, 2012). Because of competing to win RttT money, these states changed how teaching practices and teacher leadership roles would be evaluated even for states that did not win RttT funds (Boser, 2012; NCTQ, 2011a).

States also became interested in reforming teacher evaluation systems to receive

waivers from targets required by the NCLB Act and subsequent AYP measures (USDOE, 2011c). The NCLB Act required that all students be proficient in reading and mathematics by the year 2014. Secretary Duncan testified to Congress in 2011 that the USDOE estimated 82% of schools could fail to meet expectations from the NCLB Act (Duncan, 2011b). Although a study from Usher (2011), with the Center on Education Policy (CEP), put the potential number of schools failing to meet AYP at about 48%, Duncan stated, "Whether it's 50%, 80% or 100% of schools being incorrectly labeled as failing, one thing is clear: No Child Left Behind is broken" (Dillon, 2011, para. 6).

As a result, the USDOE (2011a) offered states flexibility from the requirements of the NCLB Act in exchange for evaluating teaching practices and teacher leadership roles in much the same manner as the RttT initiative outlined–measuring the success of a teacher based on classroom performance and student growth:

> Each state that receives ESEA flexibility will set basic guidelines for teacher and principal evaluation and support systems. The state and its districts will develop these systems with input from teachers and principals and will assess their performance based on multiple valid measures, including student progress over time and multiple measures of professional practice, and will use these systems to provide clear feedback to teachers on how to improve instruction. (p. 1)

Student growth on standardized tests was designed to determine the effectiveness of a teacher, and the growth of students under the tutelage of a teacher was implemented as an important factor in making personnel decisions including continued employment for teachers (Braun et al., 2010; Hull, 2011; Lohman, 2010). The RttT initiative mandated that Local Education Agencies (LEAs) "establish clear approaches to measuring student growth" and design "rigorous, transparent, and fair evaluation systems for teachers . . .

that . . . take into account data on student growth" (USDOE, 2009, p. 9).

As Fuhrman and Elmore (2004) noted, "The focus on performance, on actual student achievement, is quite a change from the way states previously gauged the health of local schools" (p. 4). Teaching practices and teacher leadership centered on performance and outcomes of the students, not necessarily the processes and inputs of teachers (Fuhrman & Elmore, 2004). Specifically, teachers would be at least partially assessed in terms of the value they add to their students' growth on standardized test scores within VAM (Hull, 2011). The value-added measurements of teacher performance has been considered significant because of the effect it has on future outcomes of present-day student learning (Chetty et al., 2011), and value-added components were incorporated by many states to constitute at least a portion of a teacher's final evaluation– in many states up to 50% (NCTQ, 2011a).

Purpose and Significance of Study

The purpose of this study was to investigate the effects of NCEES on teaching practices and teacher leadership in six local high schools. The topic displays a high rate of present-day relevancy as the RttT initiative has spurred unprecedented changes in teacher evaluation systems (NCTQ, 2011a). Thirty-two states have formally modified their teacher evaluation systems in response to the RttT initiative, including the requirement of annual teacher evaluations (NCTQ, 2011a). Using student-learning outcomes as a measure within a teacher's evaluation, including student growth being a preponderant criterion and the ability to professionally develop teachers based on teacher evaluation results, has been considered the centerpiece of the RttT initiative (NCTQ, 2011a). These states, one of which is North Carolina, will need to assess their newly implemented evaluation systems in order to ascertain their viability, which was the aim of

this research study of NCEES.

Important topics need to be explored in relation to NCEES, such as whether teachers are observed by their administrators consistently and given meaningful feedback, whether teachers are improving teaching practices, whether teachers are seeking out ways to become leaders in their classrooms and schools, whether observations and evaluations of teachers consistently reflect a quality assessment of teaching practices and leadership roles, and whether teachers perceive the processes of NCEES to be fair and helpful. The goal of this investigation of the new teacher evaluation system in North Carolina is to better the system's ability to improve teaching practices and teacher leadership by identifying any deficiencies, communicating to stakeholders about the deficiencies, and charting out a course with stakeholders about how to improve these deficiencies. An important component of investigating the effects of NCEES would also be to uncover what the system is doing well and build upon this work. This approach is a solution-based response to the "encapsulation" description, lack of classroom observations and teacher evaluations, and "buffer" notions that Sarason (1996), Schmoker (2006), and Elmore (2000), respectively, described.

The significance of this study has two underlying features. First, it has been shown that teacher evaluation systems have been ineffective, perfunctory pencil-and-paper exercises (Danielson, 2010/2011; Weisberg et al., 2009). In moving forward, this study aimed to find out how NCEES has impacted teaching practices and teacher leadership in contrast to past findings. Successful teacher evaluation systems have been shown to cultivate teacher performance and support student achievement by "enhancing teachers professional performance and fulfillment" (Ribas, 2002, p. 5).

Second, this research is significant because of the possible benefits perceived for

students when teachers are observed and evaluated properly. Plausibly, students should achieve at higher academic levels if teachers are observed and evaluated within a teacher evaluation system that helps teachers reflect upon their teaching practices and teacher leadership and improve upon them (Danielson, 2010/2011; McREL et al., 2012; Taylor & Tyler, 2012).

Definition of Terms

Case study. A case study is "an empirical inquiry that investigates a contemporary phenomenon in depth and within its real-life context, especially when the boundaries between phenomenon and context are not clearly evidence" (Yin, 2009, p. 18). A multiple case study is a case study that involves more than one unit under inspection.

Grounded theory. Based on the work of Charmaz (2006), grounded theory involves "flexible guidelines for collecting and analyzing qualitative data to construct theories 'grounded' in the data themselves" (p. 2).

Classroom observation. Classroom observation involves an observer collecting descriptive data based on the performance of a teacher and recording data using a predetermined instrument or rubric (Danielson, 2007; McGreal, 1983; Peterson, 2000; TNTP, 2010). The results have historically been used for formative evaluation (McGreal, 1983).

Teacher evaluation. Teacher evaluation is a cumulative, periodic judgment by an administrator that describes a teacher's performance over the course of a time period and can be used for formative and summative evaluation (Darling-Hammond, Wise, & Pease, 1983; Peterson, 2000).

Teacher leadership. Teacher leadership refers to a disposition of teachers

grounded in respect and gained by mentoring teachers and modeling behaviors and practices in a school-wide manner that advances student learning (Teacher Leadership Exploratory Consortium [TLEC], 2011). A defining attribute of teacher leaders is the respect they receive not because of a position or title but by their work. For example, teachers can be leaders when they actively seek to improve school-related outcomes and their profession by designing, leading, and participating in professional learning communities (PLCs), the School Improvement Team (SIT), and other professional growth opportunities (McREL et al., 2012).

Teaching practices. Teaching practices are the instructional methods that teachers choose to facilitate learning among their students, including instructional materials and a "wide range of techniques including information and communication technology, learning styles, and differentiated instruction" (McREL et al., 2012, p. 11).

VAM. VAM measures the growth of students between two points in time gauged by results on standardized tests (Kupermintz, 2003). Growth outcomes from VAM are used to estimate the effectiveness of teachers (McCaffrey & Lockwood, 2008). VAM produces quantitative data that can be used for diagnostic purposes and to rank order teachers based on their effectiveness (McCaffrey & Lockwood, 2008).

Validity. Validity is a measure of an instrument's ability to measure what it purports to measure (Cohen, Manion, & Morrison, 2007). Sartain, Stoelinga, and Krone (2011) referred to validity as a measure of how classroom observation ratings for teachers correlate with the achievement of their students. Sartain et al. and Haertel (2013) viewed the validity of VAM as its ability to measure student growth attributable to particular teachers.

Face validity. Gay, Mills, and Airasian (2009) considered face validity

ambiguous compared to content validity. Face validity is defined as the ability of a data collection instrument to measure what it claims to measure but without the formal procedures in content and construct validity (Gay et al., 2009). Gay et al. considered face validity as an important initial screening procedure in selecting the content of data collection instruments. Gay et al. also recommended that a formal content validation process should follow face validity efforts.

Content validity. Content validity is the ability of a data collection instrument to measure the content of a subject, curriculum, or process (Gay et al., 2009).

Construct validity. Gay et al. (2009) saw construct as the most important form of validity. Construct validity is the ability of a data collection instrument to measure a hypothetical construct with the word "construct" being "nonobservable traits, such as intelligence, anxiety, and honesty, 'invented' to explain behavior" (Gay et al., 2009, p. 157). In this research project, such constructs as teaching practices, teacher leadership, and teacher evaluation systems were used.

Item validity. Gay et al. (2009) defined item validity as the ability of the items on a data collection instrument to measure what the items were intended to measure.

Sampling validity. Gay et al. (2009) defined sampling validity as how well the items on a data collection instrument samples the "total content area" (p. 155) of the construct.

Reliability. Reliability is defined by Cohen et al. (2007) as the dependability, consistency, and replicability of quantitative data over time and is concerned with precision or accuracy. Cohen et al. provided evidence that the definition of reliability in qualitative data has been contested but is related to the reliability of quantitative data. This research project borrows Cohen et al.'s ideas that the reliability of qualitative data

has close ties to that of the reliability of quantitative data–the focus will be on coding qualitative data for dependability, consistency, and replicability. The reliability of observer ratings is defined as the amount of internal consistency (inter-rater reliability) among the same observer or multiple observers (Sartain et al., 2011), while the reliability of VAM is measured based on the consistency of year-to-year correlations (Haertel, 2013) or what Cohen et al. called "stability."

Unintended consequences. Unintended consequences are side effects of newly implemented teacher evaluation systems that were unknown before implementation (Amrein-Beardsley & Collins, 2012). Considerable attention to unintended consequences has been given to using VAM within teacher evaluation systems and will be explored within the primary and secondary research of this project (Amrein-Beardsley & Collins, 2012; Collins, 2014; Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012).

Formative evaluation. Popham (2013) defined formative evaluation as an "endeavor intended to supply information that can improve a teacher's effectiveness" (p. 20). Formative evaluation is used to improve teaching practices and teacher leadership during or close to an instructional time period; or as Berman (2003) stated, getting feedback to teachers "early" and "often" (p. 7).

Summative evaluation. Popham (2013) defined summative evaluation as providing "information associated with negative decisions about a teacher, such as the denial of tenure and termination" or "positive decisions, such as recognition and financial awards" (p. 20). Peterson (2000) related summative evaluation as occurring at the end of an appointed time period and used for accountability to assure the public that teachers are competent.

Research Questions

Although there is plentiful research recorded in broader terms of teacher evaluation systems, research in North Carolina is lacking. Specifically, more work is needed to ascertain how NCEES has affected teaching practices and teacher leadership roles in North Carolina. Little research has been done on collecting teacher perceptions of NCEES. The following research questions guided the study and acted as exploratory points.

1. How has NCEES affected teaching practices in six local high schools in the county under study?
2. How has NCEES affected teacher leadership in six local high schools in the county under study?

Situation of the Researcher in the Study

The researcher is a veteran high school science teacher. The researcher has experienced classroom observations and evaluation under multiple teacher evaluation systems in two states. The researcher has been trained using NCEES instruments alongside administrators in countywide professional development. The researcher has also used NCEES instruments in the capacity of a peer observer both informally and formally. The researcher did not interview administrators who were in the process of carrying out personal classroom observations and evaluations under NCEES at the present time, nor did the researcher interview administrators who had ever carried out personal classroom observations and evaluation using NCEES or any other teacher evaluation system.

Philosophical Rationale

This project was designed using Yin's (2014) single case study approach that used embedded units (six high schools). This approach was chosen based on the research

questions and the strength of case studies to explore social phenomena at individual, organizational, and political levels (Yin, 2014). When observers engage in observing and evaluating teachers, they act in the interest of the public at the political level and in the interest of schools at the organizational level, all of which involve highly social processes (Sartain et al., 2011). The single case study, as described by Yin (2014), was a top-down research approach using the boundaries presented in the research questions of the literature review.

This case study used a mixed-methods research design (QUAL-Quan). Initial categories were synthesized using a grounded theory analysis of qualitative data from pilot focus groups with teachers and from interviews with administrators. The initial categories were refined in the literature review and used to frame questions for structured items in a questionnaire used within a survey process. Unstructured items on the questionnaire used within the primary research provided qualitative data that were coded using grounded theory analysis. Administrators for each site were also interviewed as part of the primary research and provided qualitative data analyzed using grounded theory.

Therefore, the research was primarily qualitative following Charmaz's (2006) principles. However, quantitative data within a Likert-based questionnaire were also collected and analyzed. Quantitative data were also analyzed interpreting teacher ratings across the distribution of teacher performance and across years to detect effects of NCEES on different fronts. Hence, this research project was a mixed-methods design.

Grounded theory was the best methodology to frame this project. Techniques in grounded theory have allowed researchers to follow the data and not "force preconceived ideas and theories directly upon our data" (Charmaz, 2006, p. 17). This bottom-up

approach considered teacher perceptions to construct reality which was necessary in building a case that uncovered the effects of NCEES on teaching practices and teacher leadership.

Creswell (2012) advised a mixed-methods design when multiple forms of data can provide a better understanding of a research question. Analysis of quantitative data allowed the researcher to work deductively to understand the effects of NCEES on teacher ratings within the case study framework. Quantitative data often describes what is happening without investigating why phenomena occur and often work using deductive logic (Charmaz, 2006; Gay et al., 2009). Qualitative data often use an inductive approach and were used in this research project to describe how NCEES has affected teaching practices and teacher leadership (Charmaz, 2006).

Taken together, the top-down research approach of the case study design and the bottom-up research approach of grounded theory provided methodology that generated valid data collection instruments used within the survey process of the primary research. Figure 1 provides a description of the methodology and rationale used to validate the data collection instruments used within the primary research.

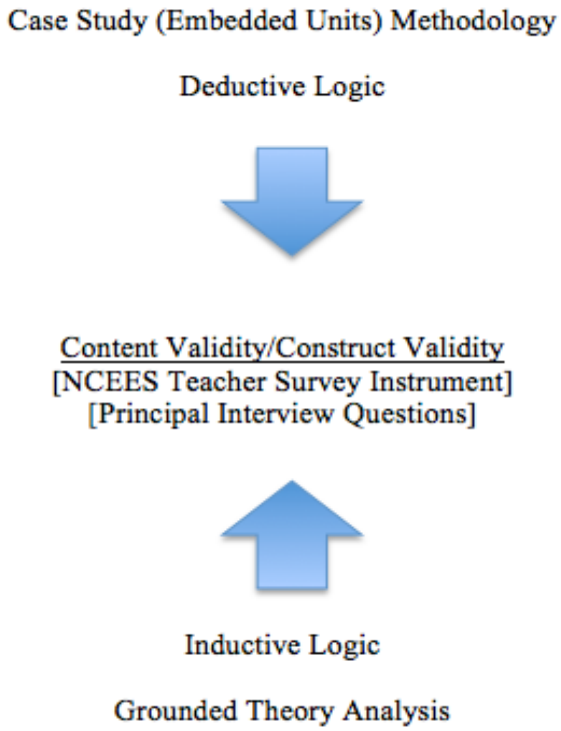


Figure 1. Methodology to Validate Data Collection Instruments.

Two instruments were used to measure the effects of NCEES on teaching practices and teacher leadership: a questionnaire that was employed within survey procedures carried out in the six high schools of the county under study and principal interview questions that were used to interview at least one administrator at each high school in the county under study. Using the top-down analytical research approach from a case study methodology and a grounded theory analysis of qualitative data from the bottom-up, all possible efforts were placed on validating the NCEES Teacher Survey and the NCEES Principal Interview questions for content and construct validity standards by Gay et al. (2009). Ultimately, the NCEES Teacher Survey (see Appendix A and Appendix B) and NCEES Principal Interview (see Appendix C) were two important

procedures (among others; see Chapter 3) used to collect data to measure the effects of NCEES on teaching practices and teacher leadership in the six high schools in the county under study.

To summarize, many authors have noted a lack of quality in teaching practices and teacher leadership of teachers in U.S. public schools (Frase & Streshly, 1994; Goodlad & Klein, 1970; Schmoker, 2006). Also, studies have noted a lack in quality of U.S. public school graduates (ACT, 2009, 2011; CCA, 2012). However, studies have also shown that teachers are widely observed and evaluated with high marks within teacher evaluation systems (Weisberg et al., 2009).

This study has been designed to measure the effects of NCEES on teaching practices and teacher leadership in order to observe whether the recent implementation of NCEES has overcome historical barriers of teacher evaluation systems. This study involved a case study approach concurrent with grounded theory analysis. Also, this study was a mixed-methods endeavor.

Chapter 2: Literature Review

Overview

Literature reviews have served to propose "knowledge about a particular field of study, including vocabulary, theories, key variables and phenomena, and its methods and history" (Randolph, 2009, p. 2). This literature review has served as a secondary research method added to the primary research. Randolph (2009) described the actual methods that the researcher carried out to collect and analyze data as the primary research methods. This literature review modeled Randolph's Taxonomy which was designed based on Cooper's (1988) Taxonomy of Literature Reviews (p. 2). A focus on research outcomes dominated this literature review (Randolph, 2009, p. 2).

The research outcomes from this secondary research provided data from previous studies that used sound methodology and provided data that added to the various quantitative and qualitative data provided by the primary research methods of this study. Randolph (2009) cited Gall, Borg, and Gall (1996) and suggested that an important component of a scientific approach to a literature review involved "seeking support for grounded theory" (p. 2), which was the qualitative method of data analysis for the primary research.

The foremost goal of this literature review was to generalize findings about teacher evaluation systems and their effects on teaching practices and teacher leadership both in present-day and historical applications in an attempt to add to the quantitative and qualitative measures provided by the primary research. This format modeled Randolph's (2009) foci and fulfilled Soy's (1997) call for historical analysis within a literature review to build a case for the primary research.

An important function of this literature review (secondary research) was to serve

as a secondary source of qualitative data to help validate findings from the primary research (Randolph, 2009). As Randolph (2009) suggested,

> If one thing must be realized about conducting and reporting a literature review it is that the stages for conducting and reporting a literature review parallel the process for conducting primary research. With a few modifications, what one knows about conducting primary research applies to conducting secondary research (i.e., a literature review). (p. 4)

In following Randolph's (2009) suggestion, within this literature review, a problem was formulated, research questions posed, data collected and evaluated, and conclusions drawn. The conclusions drawn from the literature review, as a secondary source of qualitative data, were added to and compared to the quantitative and qualitative data supplied by the primary research methods. The research outcomes of this literature review have served as a data source for the eventual triangulation of data in the primary research methodology.

Design of literature review. The proceeding literature review also followed suggestions from Charmaz (2006). The literature review was delayed and completed alongside collecting qualitative pilot data involving focus groups with teachers and interviews with principals, in keeping with Charmaz's "delayed literature review" (pp. 6, 12). The pilot data were categorized based on initial and focused coding of the collected qualitative data. Categories were developed using Charmaz's "constant comparative methods" and following Charmaz's format for writing initial and advanced memos. These categories were then used to help develop research questions for the literature review and perform further investigation of the literature surrounding the collected qualitative data. These categories served as a basis to develop Charmaz's "theoretical

saturation" of categories (p. 96).

Charmaz's (2006) "delayed literature review" was also in keeping with using the literature review as a source of secondary research that provided support to the primary research, which Randolph (2009) suggested. Throughout Charmaz's work, as the author collected qualitative interview data concerning systemically ill patients, she investigated and referred to the medical literature pertaining to her particular research. Charmaz's approach mirrored the methodology of this literature review in that, while collecting and analyzing qualitative pilot data, the literature was reviewed for pertinent leads.

Research questions for literature review. The research questions that guided this literature review were as follows:

1. What have been the perceived limitations of teacher evaluation systems historically and leading to the present day?
2. How have the processes of teacher evaluation systems affected teaching practices and teacher leadership roles historically and leading to the present day?

The research questions were synthesized from data collected from pilot focus groups and administrator interviews (using Charmaz's [2006] categorical coding scheme) and after surveying existing literature surrounding teacher observation and evaluation.

The two research questions were not exclusively driven and independent of one another. Throughout this literature review, these questions were approached in a collective fashion. For instance, in researching the quantity of evaluations received by teachers annually, the quantity of teacher observations and evaluation that teachers may (or may not) have received was researched in an attempt to investigate Research Question 1. However, during this same juncture of research, a lack or abundance of teacher

observations and evaluation experienced by teachers naturally flowed into answering Research Question 2. Research Questions 1 and 2 may be thought of as interwoven research questions; and, in many cases, while investigating one, knowledge of the other naturally appeared.

Elucidating terms from research questions. In Research Questions 1 and 2, the term "historically" loosely referred to the era in which teacher evaluation was popularized, which was in the 1970s (Shinkfield & Stufflebeam, 1995). Such an historical investigation within a literature review was applicable within the bounds of case study research. Soy (1997) pointed out that the subject in a case study "is likely to be intricately connected to . . . historical issues . . . providing wide ranging possibilities for questions and adding complexity to the case" (online document, para. 5). Creswell (2003) also advised to set proper boundaries of a case study by identifying the time period and places in which the processes involved were under investigation.

In Research Question 2, the processes involving teacher evaluation systems referred to the processes that occurred within teacher evaluation systems across states, not the structural design of state and local teacher evaluation systems. To analyze the specific design and structure of teacher evaluation systems would demand a broad examination of federal, state, and local legislative mandates, which was out of the scope of this study. Also, such an investigation would not be advantageous in answering the research questions.

In Research Question 2, "teaching practices" refers to the components of a teacher's performance in relaying knowledge to their students. Specifically, "teaching practices" refers to the instructional procedures purposely chosen by a teacher to relay information to their students and can range from a teacher's lesson plan to how a teacher

chooses to answer a question from a student during class. Other informal, perhaps more subtle examples of "teaching practices," would involve teachers reflecting on their instructional practices in order to improve the achievement of their assigned students or analyzing data produced by their assigned students from tests and other classroom assessments.

The title "teacher leadership" in Research Question 2 referred to teachers who were assigned, or voluntarily pursued, roles other than those naturally carried out within their classrooms (with their assigned students) in order to increase school-wide student achievement. Examples of "teacher leadership" could range from a teacher mentoring peer teachers, acting as a mentor to a struggling teacher, or designing and leading a PLC. The title "teacher leadership" has been treated as any teacher-led activity where the achievement of their assigned students did not directly correlate with the activity, but the activity served to increase student achievement by leading and inspiring other teachers to improve student achievement.

While defining "teaching practices" was a relatively simple task, defining "teacher leadership" was not. This definition was not meant to be exhaustive but broadly based for the needs of this study. The Institute for Educational Leadership (IEL, 2001) admitted that the "literature on the teacher as a leader is thin" and broadly referred to teachers who were leaders as "teachers who seek and find challenge and growth" (p. 4). The authors also said that teachers who were leaders made "their presence felt beyond the classroom walls" (IEL, 2001, p. 4).

The *Teacher Leader Model Standards*, composed by the Teacher Leadership Exploratory Consortium (TLEC, 2011), were released in Washington, D.C., and in them the Consortium commented on what makes a teacher a leader:

> The term "school leader" often refers to the principal, director of curriculum, pupil services director, or other building- or district level administrators. Teachers who serve in leadership roles may do so formally or informally. Rather than having positional authority, teachers become leaders in their schools by being respected by their peers, being continuous learners, being approachable, and using group skills and influence to improve the educational practice of their peers. (p. 11)

The terms "formally" or "informally" and the list of leadership activities from TLEC's (2011) definition coincided with the definitions used in this study. TLEC described "teacher leadership" as teachers who were assigned ("formally") or voluntarily pursued ("informally") roles, other than those naturally carried out within their classrooms with their assigned students (i.e.,"being respected by their peers, being continuous learners, being approachable, and using group skills and influence to improve the educational practice of their peers," (p. 11). It was also important to note that the leadership activities listed by TLEC's definition pertain to activities that were generally not carried out within a classroom with a teacher's assigned students; rather, the activities of "teacher leadership" worked to improve school-wide student achievement.

Perhaps the most important resource that defined "teacher leadership" was the NCEES rubric provided in the handbook entitled the *North Carolina Teacher Evaluation Process* (NCTEP) (McREL et al., 2012). The NCEES rubric has been used to rate North Carolina teachers and will be the document used throughout this study for the purpose of determining the expectations of North Carolina teachers.

The NCEES rubric guided the definition of "teacher leadership" used in this study by defining what teachers as leaders do in the classroom (Element A), in the school

(Element B), for the teaching profession (Element C), and for schools and students (Element D) (McREL et al., 2012). With the exception of Element A (what teachers as leaders do in the classroom), the tasks listed for Elements b-d reinforced the idea that "teacher leadership" involved teacher-led activities that were assigned or voluntarily pursued that improved school-wide student achievement. The tasks from NCTEP Elements b-d included example teacher activities such as attending professional learning community meetings, being aware of the goals in the school improvement plan, taking advantage of professional growth opportunities, knowing about the policies and practices affecting student learning, and understanding the importance of ethical behavior (McREL et al., 2012).

The terms "teacher evaluation" and "teacher observation" also required delineation. The term "teacher observation" has been defined as an individual in an administrative or peer position that judges the performance of a teacher within the classroom setting (McGreal, 1983; Peterson, 2000). "Teacher evaluation" has referred to the macroscopic process of a teacher being judged and rated based on their job performance which can include classroom observations, portfolio work, self-reflective pieces, student performance in VAM, and an assessment of a teacher's commitment to their work outside of the classroom yet in the environment of schooling (teacher leadership) (Peterson, 2000).

Processes Described Within Historic Ineffective Teacher Evaluation Systems

Many authors have engaged in large-scale investigations of teacher evaluation systems and agreed that there are two major reasons for teacher evaluation–to ensure quality control measures among teachers and to provide teachers with feedback in order to improve their performance (Danielson, 2010/2011; Darling-Hammond, 2010; Levin,

1979; McGreal, 1983; Wise, Darling-Hammond, McLaughlin, & Bernstein, 1984). Although Griffith (1973) found that evaluating teachers was difficult work, time-consuming, generally disliked by teachers and administrators, and historically teacher evaluation had not been required, Griffith attested that evaluating teachers was important because it gave stakeholders information about what was going on in the classroom and encouraged and assisted teachers to improve. Whether because of difficulty or other circumstances, teacher evaluation systems have not historically garnered attention and have been deemed an inconsistent and ineffective exercise in need of repair (Elmore, 2000; Griffith, 1973; Harris, 1986; Medley & Coker, 1987; Schmoker, 2006; Soar, Medley, & Coker, 1983; Weisberg et al., 2009; Wise et al., 1984).

Ineffective teacher evaluation systems have been consistently noted in the literature dating back to the 1970s and 1980s (Griffith, 1973; Soar et al., 1983; Wise et al., 1984). However, the RttT initiative has recently served as a major impetus for revamping teacher evaluation systems including the requirement that veteran teachers be evaluated at least annually. The RttT initiative also focused on formative evaluation using growth rubrics rather than checklists and subjective rating designs.

Alongside a focus on formative evaluation, the RttT initiative made student growth a preponderant criterion on which teacher evaluation was to be based, while making it easier to dismiss incompetent teachers (NCTQ, 2011a). Although there have been unprecedented changes in teacher evaluation systems recently, revamped teacher evaluation systems have been under scrutiny in order to gauge their validity and reliability (Amrein-Beardsley & Collins, 2012; Sartain et al., 2011; Tennessee Department of Education [TNDOE], 2012).

Teacher evaluation systems have been criticized historically because of their

dependency on observing teacher behavior during the complex teaching and learning process and exclusively measuring and judging the input of teachers without being concerned with the output of the learning process, specifically how much students learn (Darling-Hammond et al., 1983; Fuhrman & Elmore, 2004; Medley & Coker, 1987). Though judging the performance of teachers solely based on their inputs has been the traditional approach, Darling-Hammond et al. (1983) pointed out that "even proponents of classroom observation recognize its limitations," namely that "observer bias, insufficient sampling of performance, and poor measurement instruments can threaten the reliability and validity of results" (p. 306).

To overcome an "insufficient sampling of performance" (Darling-Hammond et al., 1983, p. 306), processes within teacher evaluation systems have ideally originated and continuously occurred with multiple observers who judge the instructional performance of teachers in action within the classroom (Danielson & McGreal, 2000; McGreal, 1983; Schmoker, 2006). The more sampling (judgments from observers), the greater the reliability of classroom observations and evaluation results within evaluation systems (MET, 2012; Peterson, 2000; Prince, 1984). The minimal level of expectations in observing and evaluating teachers in action has historically not been met–authors have noted that many teachers, especially veteran teachers, teach for many years without being observed or evaluated within evaluation systems (Griffith, 1973; Schmoker, 2006; Weisberg et al., 2009; Wise et al., 1984).

Teachers have expressed a desire to be observed within evaluation systems for the sake of improving their performance (Bill and Melinda Gates Foundation, 2012; Harris, 1986); and in the past, teachers have voluntarily requested to be observed within evaluation systems in order to improve (Griffith, 1973). The desire to be regularly

observed and evaluated has been expressed more consistently by younger teachers according to the American Institutes for Research (AIR) and American Federation of Teachers (AFT) (2011), which carried out an investigation using focus groups, survey reviews, and case studies. The AIR and AFT (2011) study found that younger teachers desired "more frequent feedback on their teaching" from principals than did veteran teachers (p. 6). In the *Retaining Teacher Talent Survey*, from the AIR and AFT, three of four beginning teachers said that they

> prefer having a principal who frequently observes my classroom and gives me detailed feedback on how I'm doing as opposed to having a principal who conducts formal observations of my teaching only once a year and gives me only general feedback. (p. 7)

This could be important for the retention of younger teachers as the AIR and AFT noted that younger teachers leave the profession (51% higher) and switch schools more often (91% higher) than their older colleagues (p. 7).

This was the first of four significant and repetitious themes throughout an examination of the literature, and from categories derived using a grounded theory, qualitative analysis of pilot data that were collected–that there has been a lack of quantity in observing and evaluating teachers within teacher evaluation systems. This finding corresponded with the observations of Darling-Hammond et al. (1983) about an "insufficient sampling of performance" (p. 306) that has existed in observing and evaluating teachers (Griffith, 1973; Peterson, 2000; Schmoker, 2006; Wise et al., 1984). Teachers often work in solitary conditions without specific feedback that could potentially improve their performance (Huffman & Hipp, 2003; Marshall, 2009).

When administrators have observed teachers in action, even in a consistent

manner that attempted to overcome Darling-Hammond et al.'s (1983) "insufficient sampling of performance," researchers have argued that the results may not have provided valid and reliable data (McIntyre, 1980; Medley & Coker, 1987). The issues surrounding the validity and reliability of observer judgment have historically denigrated the quality of teacher evaluation systems (MET, 2012).

McIntyre (1980) and Medley and Coker (1987) submitted that although teachers may be consistently observed, there may be significant bias within the structure of evaluation systems and the results were not reliable. Both McIntyre and Medley and Coker blamed observation effects such as an "observer effect" (p. 36) and a "halo effect" (p. 242), respectively, which hindered evaluation systems in providing an unbiased judgment of teaching practices and teacher leadership roles and, in turn, provided low quality judgments which affected the quality of the evaluation of teachers.

MET (2012) found that the same teacher was "often rated differently depending on who did the observation and which lesson was being observed," despite the training that "some 900 trained raters" received (p. 5). MET (2012) concluded that students provided a more reliable and predictive source of teacher evaluation through the use of a student survey than the trained raters when comparing results of the raters and outcomes from student surveys to student gains on state tests–and this conclusion was derived after averaging scores from multiple observations from different raters in order to increase inter-rater reliability (p. 5).

Another possible issue affecting the quality of teacher observations has been the observation checklists, rating approaches, and observation rubrics within teacher evaluation systems that historically have been used to judge teachers in action (Guarino & Stacy, 2012; Soar et al., 1983). Teacher observation instruments used to judge the

effectiveness of teachers and their teaching practices and teacher leadership roles have evolved from checklists used in the 1970s and 1980s into standards-based growth rubrics (Danielson, 2007; Griffith, 1973; McGreal, 1983; Wise et al., 1984).

However, authors have cautioned the overreliance of using instruments to judge the performance of teachers during classroom observations based on the complexity of teaching as an art rather than labor (Darling-Hammond et al., 1983; Peterson 2000). Observation instruments used to judge teachers presented what McGreal (1983) termed "a narrowed focus on teaching" (p. 70). While narrowing the focus to actions that were thought to improve student achievement was necessary to standardize judgments about how teachers were expected to perform (i.e., Flander's System, Hunter's Seven Components of Teaching, Charlotte Danielson's Framework), authors argued that by reducing the expectations of teaching to expected behaviors, observers may have been judging teacher behaviors that "lack the minimum properties necessary for accurately measuring the performance of teachers" (Soar et al., 1983, p. 243).

An overreliance on instruments to judge the performance of teachers has been shown to narrow teaching practices and teacher leadership in order to fulfill the elements on the instrument from which they were being judged, regardless of the quality within the instrument (Milner, 1991). This situation would have been ideal if instruments used to judge teaching practices and teacher leadership roles were highly correlated with student learning. However, Soar et al. (1983) found that low inference and broadly based measuring scales (purportedly more objective than high-inference and narrowly based instruments) used by a "reasonably sophisticated observer," "had no validity as predictors of teacher effectiveness" (p. 244).

This has been a second significant and repetition theme in the literature, and from

categories derived from a grounded theory, qualitative analysis of pilot data that were collected–teacher evaluation systems have provided low quality observations and evaluation of teachers for two primary reasons. First, observers have shown to be inaccurate raters of teacher performance, at least compared to other sources of data (McIntyre, 1980; Medley & Coker, 1987; MET, 2012); and second, observers may have been relying on a set of assumed teacher practices within observation rubrics that may not have correlated with practices that improved student achievement (Guarino & Stacy, 2012; Milner, 1991; Soar et al., 1984).

Authors have attributed the shortcomings of teacher evaluation involving classroom observation techniques that rely solely on administrator judgment to the complexity of the teaching and learning process (Peterson, 2000). Researchers have cautioned that it is has been unclear which teacher behaviors are responsible for student learning (Marshall, 2009; Milner, 1991; Peterson, 2000; Shinkfield & Stufflebeam, 1995). Thompson (1962) recognized the complexity of teaching many years ago:

> It is a relatively simple matter to measure the productive output of factory and clerical workers, but once the yardstick is applied to an individual whose position involves mental processes, the matter becomes quite difficult. Take the teaching profession, for example. Many devices have been evolved to evaluate the teacher, yet no one method has proved entirely satisfactory. Why? There are too many intangibles involved. (p. 169)

In measuring the effectiveness of teaching practices through teacher observations and evaluation, an approach has recently been employed to overcome the difficulty of measuring and judging the "inputs" of teachers (Daley & Kim, 2010; Furhman & Elmore, 2004). Rather than concentrating on how teachers perform within teacher observations

and evaluation (inputs), teacher evaluation systems have recently incorporated a new measure of teaching practices and teacher leadership that has analyzed student-learning outcomes (outputs) (Fuhrman & Elmore, 2004; NCTQ 2011a).

VAM has been a newly implemented measure of student learning in recently redesigned teacher evaluation systems (Amrein-Beardsley & Collins, 2012; Daley & Kim, 2010; NCTQ, 2011a; Sartain et al., 2011; TNDOE, 2012). This approach has measured student growth on standardized test scores from two points in time in an attempt to describe the value added or lost by a teacher (Braun et al., 2010; Cody et al., 2010).

However, difficulties in the use of VAM methods have shown to produce instability in teacher ratings from year to year (Baker et al., 2010), and teachers have expressed discontent with the lack of transparency in the sophisticated statistical methodology that VAM has employed (Amrein-Beardsley & Collins, 2012; National Academy Press [NAP], 2009). For these reasons, researchers have advised policymakers to use caution in employing VAM outcomes for high-stakes decisions and have advised that VAM be used in conjunction with multiple measures of teacher evaluation (Darling-Hammond, Amrein-Beardsley et al., 2012; Lockwood, Louis, & McCaffrey, 2002; McCaffrey et al., 2004).

VAM embodied the reform efforts of the RttT initiative in that it has used student growth measures as a preponderant criterion within teacher evaluation systems (NCTQ, 2011a). Although VAM enabled the evaluation of teachers to be solely reliant upon teaching and learning outcomes without the concern of a lack of quantity or quality from classroom observations, Amrein-Beardsley and Collins (2012) revealed unintended consequences that stemmed from the implementation of VAM within teacher evaluation

systems in the Houston Independent School District (HISD). Teachers within the Amrein-Beardsley and Collins study expressed confusion about the outcomes of VAM and observed that VAM outcomes oftentimes differed from the judgments of observers. This type of discrepancy could threaten the validity of VAM in the future (Amrein-Beardsley & Collins, 2012; TNDOE, 2012).

However, because VAM potentially has overcome observer bias in collecting qualitative data about teacher performance (Goe, 2008) and relies on using quantitative data that have been designed to estimate teacher effects on student outcomes, quantitative VAM outcomes have been analyzed alongside qualitative judgments about teacher performance from classroom observers (TNDOE, 2012). This has provided a way for researchers at the federal and state levels to measure the validity of qualitative judgments from classroom observations compared to quantitative measures of VAM (TNDOE, 2012).

Regardless of the advantages of employing VAM, opponents have been concerned about the overreliance on standardized testing that has been necessary to evaluate teachers, and that standardized testing has been being misused to evaluate teaching practices and teacher leadership (Harris, Smith, & Harris, 2011; Ravitch, 2010). A concern in the literature has surrounded using standardized tests for high-stakes accountability purposes and whether the outcomes of those tests reflect what students memorized for the test rather than measuring the quality of teaching practices and teacher leadership (Koretz, 2008).

An important factor considered in the literature was whether the performance of a student on a standardized test could be attributed to the quality of teaching practices and teacher leadership roles or were there other mitigating variables that affected outcomes

such as the collaboration among students within the class, parental involvement, and the competence of the principal (Baker et al., 2010). Opponents of VAM have claimed that relying upon test scores from standardized tests to reward or punish teachers and principals may cause unintended consequences such as teaching to the test; narrowing the curriculum; negating collaboration among teachers; gaming the system; and cheating by teachers, administrators, and other stakeholders (Koretz, 2008; Ravitch, 2010).

The use of VAM to evaluate teachers represented a third significant and repetitious theme uncovered from the literature and from categories derived from a grounded theory, qualitative analysis of pilot data that were collected. First, authors cautioned the use of VAM based on the complexity of measuring and "disentangling" teacher effects from "the many influences on student progress," which could give rise to the instability of teacher ratings (Darling-Hammond, Cook, Jaquith, & Hamilton, 2012, p. iii). Second, because the methodology involved in VAM has been sophisticated beyond the understanding of teachers and other stakeholders and it may not be directly apparent how VAM has helped to increase student learning, it has been argued that VAM has lacked transparency and is punitive in its approach to holding teachers accountable for student learning (Amrein-Beardsley & Collins, 2012; NAP, 2009; Ravitch, 2010). The third concern centered on unintended consequences of VAM on teachers and administrators, specifically centered on whether this novel reform effort has affected teachers and administrators in ways that could be considered detrimental to the teaching and learning process (Amrein-Beardsley & Collins, 2012; Koretz, 2008).

After collecting evaluation data for teachers in terms of teaching practices and teacher leadership, authors have argued that the outcomes have not been used to spur improvement–the outcomes have been exclusively used to judge the minimal

competencies of teachers rather than to develop them (formative evaluation) or dismiss them (summative evaluation) (Braskamp, 1980; Danielson & McGreal, 2000; Weisberg et al., 2009). Braskamp (1980) and Danielson and McGreal (2000) advocated that teacher evaluation outcomes could be successfully employed in a hybrid manner to help teachers improve and to dismiss ineffective teachers. However, outcomes from teacher evaluation systems have rarely been used in such fashion (Daley & Kim, 2010). Braskamp viewed this paradox by considering the need of professional autonomy for teachers as decision makers but also providing a respectful critique of their work. Braskamp viewed this inherent dilemma to both formatively help teachers improve and summarily judge teachers for the sake of quality control as a potential "dysfunction" of teacher evaluation systems (p. 98). The enduring paradox has endured to this day–how do administrators observe and evaluate teachers to help them improve while maintaining trusting relationships among teachers so they are not fearful that administrators are observing and evaluating them to potentially dismiss them (Peterson, 2000)?

There has been an upshift in using data from teacher evaluations to improve the teaching practices and teacher leadership roles of teachers rather than to simply check for minimal competencies among teachers (Danielson, 2007); however, advocates recently advised administrators to make more summative decisions to dismiss ineffective teachers on the basis of outcomes stemming from evaluation data (Weisberg et al., 2009). In places such as Washington D.C., Houston, TX, and Central Falls, RI, recently there have been unprecedented dismissals of teachers based on summative judgments stemming from teacher evaluation data (Amrein-Beardsley & Collins, 2012; Kaye, 2010; Turque, 2010).

This tug-of-war paradox of using teacher observation and evaluation outcomes to

improve teachers or dismiss them has been noted consistently in the literature (Braskamp, 1980; Danielson & McGreal, 2000; McGreal, 1983; McLaughlin & Pfeifer, 1988). Danielson and McGreal (2000) noted the paradox has been generated between legislators and policymakers who demand "quality assurance and accountability" and educators who think that "teacher evaluation should be designed for the purpose of professional development and the improvement of teaching" (pp. 8-9).

There has been an upshift in the use of teacher observations and evaluation formatively to improve teachers (Daley & Kim, 2010) and in a summative manner to dismiss teachers when necessary (NCTQ, 2011a). This paradox presented the fourth significant and repetitious theme in the literature and from categories derived from a grounded theory, qualitative analysis of pilot data that were collected. Historically, outcomes from teacher observation and evaluation have been used sparingly for any purpose; and as a result, outcomes from teacher observations and evaluation have been largely viewed as perfunctory (NCTQ, 2011a; Peterson, 2000; Sartain et al., 2011; Wise et al., 1984).

VAM has introduced a new innovative form of teacher evaluation that has provided an evaluation of teachers independent of the judgments made by administrators and other observers (Goe, 2008), and VAM has added another element to multiple measures used in summative evaluation (Hanover Research, 2012). By coupling the quantitative growth data from VAM with qualitative classroom judgments from administrators, a new era of teacher evaluation has been conceived that has incorporated an objective component of teacher evaluation (Hanover Research, 2012).

Although recent reform efforts to overhaul teacher evaluation systems have paved the way for outcomes of teacher evaluation to be used in a formative or summative

manner, there still exists the contentious debate over whether outcomes should focus more on formative or summative purposes or whether outcomes can be used successfully in a hybrid manner (Daley & Kim, 2010; NCTQ, 2011a). This fourth significant and repetitious theme centered on how outcomes from teacher observations and evaluation should be used and concentrated on the paradox of using observations and evaluation outcomes to improve teachers, to dismiss them, or both. This fourth and final theme stemmed from reviewing the literature and from categories derived from a grounded theory, qualitative analysis of pilot data that were collected.

To summarize, there have been four significant and repetitious categories from the pilot data that were found relevant and worthy to review in the literature. The categories represented the most significant and repetitious ideas surrounding ineffective teacher evaluation systems as presented in the pilot data and literature. These categories will be used as a basis for structuring the remainder of this literature review.

Unbeknownst before collecting and analyzing pilot data and reviewing the literature, these four significant and repetitious categories displayed flow and showed dependency upon one another. This observation became apparent when applying Charmaz's (2006) axial coding of qualitative data while investigating their relatedness.

The pilot data were designed to provide a body of data to investigate pertinent leads for guidance in completing the literature review. There were two focus groups of six teachers and two administrator interviews. The questions were designed purposely to be as open-ended as possible in order to give teachers and administrators the opportunity to guide discussions.

There were about 2 hours of transcribed material from recorded discussions from teachers and administrators. The qualitative data were then initially coded line-by-line

followed by a more focused coding to generate categories and finally analyzed using axial coding to describe an overall picture of the conditions, actions, and consequences involved in the effects of NCEES (Charmaz, 2006).

These categories represented areas that will be investigated further in the primary research methods of this research project in Chapter 3. This flow and dependency has been illustrated in Figure 2.

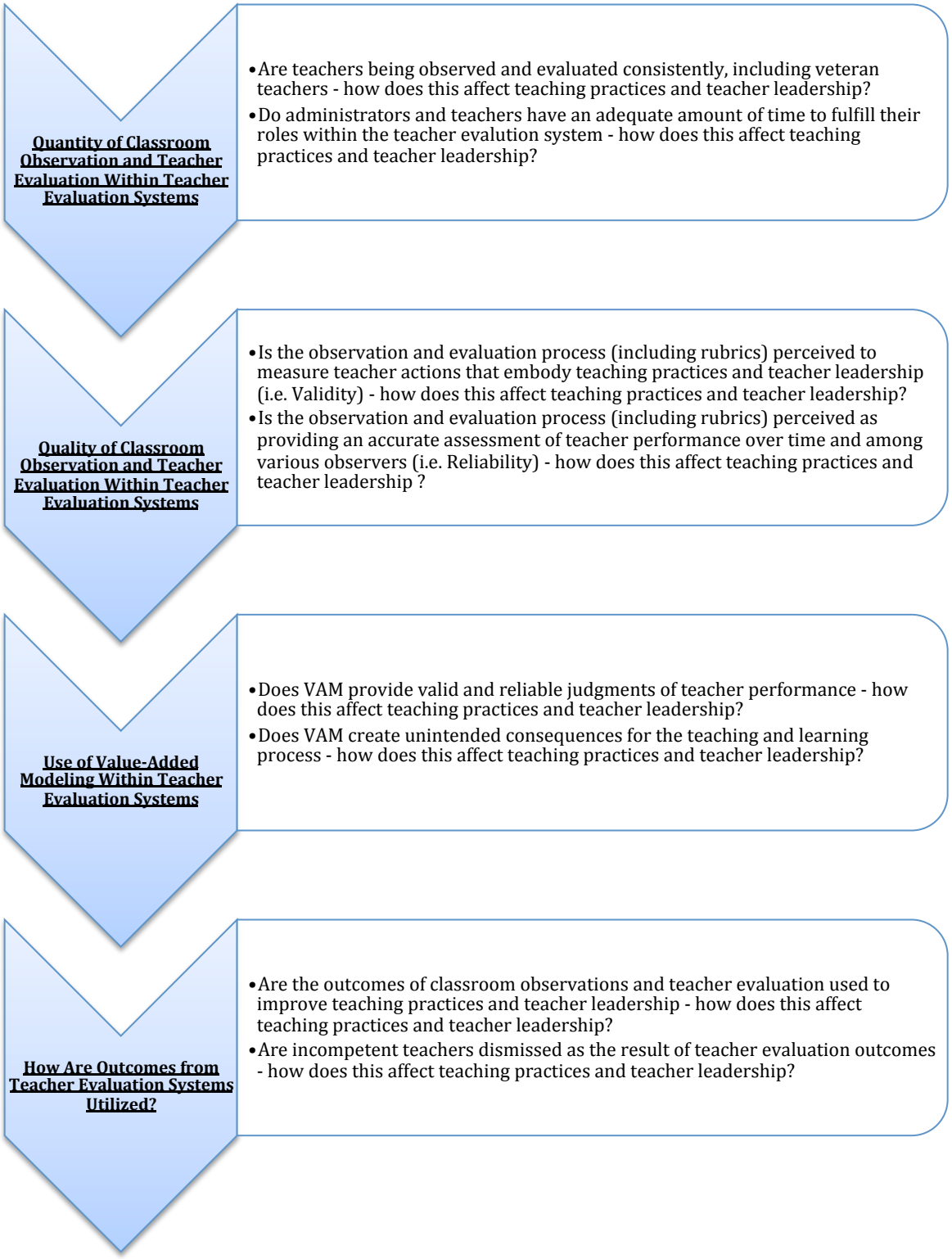


Figure 2. Four Grounded Theory Categories from Pilot Data–Refined by the Literature.

The Effects of the Quantity of Classroom Observations and Teacher Evaluation

Consistency of observation and evaluation processes. Teacher evaluation has existed in some form since there have been teachers, probably dating back to the pupils of Socrates (Shinkfield & Stufflebeam, 1995). Teacher evaluation was popularized in the 1970s, picked up momentum in the 1980s, and eventually took center stage in the 1990s and beyond (Shinkfield & Stufflebeam, 1995). According to Shinkfield and Stufflebeam (1995), "The last 15 years has seen a dramatic development of the technology of teacher evaluation" (p. 1).

As early as the 1960s, it was shown that teacher evaluation systems were lacking consistency in collecting data about the performance of teachers, and the negative effects on teaching practices and teacher leadership were apparent. Griffith (1973) reported in the 1960s that administrators visited classrooms infrequently. It was sometimes spans of years that passed before teachers were observed and evaluated; and as a result, those evaluations were "perfunctory and valueless" (Griffith, 1973, p. 3).

Griffith (1973) cited a 1965 study by the NEA in which 20% of secondary school probationary teachers received no evaluation at all for 1 year preceding the study; 42% of secondary school tenured teachers received no evaluation for 1 year preceding the study; and about 11% of the teachers observed had to request that administrators observe them in the classroom environment (pp. 3-4).

An NCTQ (2011b) survey of teachers (n=1,317) measuring the quality of teacher evaluation systems showed that about 25% of teachers did not receive any feedback within evaluation systems, about 25% received feedback once a year, about 25% received feedback at least twice a year, about 20% received monthly feedback, and about 5% received weekly feedback. Although about 25% of teachers reported they received no

feedback within teacher evaluation systems, the NCTQ's (2011b) survey showed that over 60% of teachers who had received feedback considered it "at least somewhat helpful" (p. 22).

Teacher evaluation systems have historically hinged on whether or not teachers were observed and evaluated within the classroom environment (Good & Brophy, 1984; McGreal, 1983). Good and Brophy (1984) expressed discontent that "In many school districts teachers are visited only once or twice a year, and many veteran teachers are not visited at all. . . . Under these circumstances, it is ridiculous to even talk about possible advantages of observation" (p. 145).

The quantity of teacher observations and evaluations within teacher evaluation systems affected the teaching practices and teacher leadership roles of teachers (Goodlad & Klein, 1970; Griffith, 1973; Schmoker, 2006; Taylor & Tyler, 2012; Wise et al., 1984). When there was an absence of teacher observations and evaluation within an evaluation system, there was a lack of improvement in teaching practices and teacher leadership roles (Goodlad & Klein, 1970; Wise et al., 1984). When there were consistent teacher observations and evaluation within an evaluation system, there was an improvement in teaching practices and teacher leadership roles (McLaughlin & Pfeifer, 1988; Schmoker, 2001; Taylor & Tyler, 2012).

Schmoker (2001) accounted for "consistent and impressive gains at every level" in a California elementary school that was spurred by consistently observing and evaluating teachers in action (throughout the 1998-1999 and 1999-2000 school years) and enacting collaboration measures among teams of teachers based on data stemming from the observations (pp. 46-47). After the consistent observation and evaluation of teachers, and collaboration among teachers, Schmoker (2001) detailed how every component for

reading outcomes on the Stanford 9 showed growth between 1998 and 2000.

In later research, Schmoker (2006) retrospectively referred to the consistent observation and evaluation of teaching practices and teacher leadership as the catalyst that drove improved teacher performance and collaboration in this California elementary school. Schmoker (2006) contrasted consistently observing and evaluating teachers and sharing the results to improve with the consistent lack of teacher observation and evaluation when he said,

> Almost every class reveals the instructional consequences of our historic failure to monitor or supervise instruction and arrange for teachers to work in teams so they can effectively teach. . . . Years of isolation–from colleagues, from constructive supervision–account for this alarming gap between what we know and do. This gap can be seen in all schools, across the socio-economic spectrum. (pp. 16-17)

Schmoker's (2006) insight can be confirmed dating back Goodlad and Klein's (1970) study. Goodlad and Klein portrayed a disheartening description of schooling in the U.S. Goodlad and Klein collected a myriad of data in which researchers investigated the organization, curriculum, nature of a school day, classroom climate, and the uniqueness of the programs of a particular school.

Goodlad and Klein (1970) was based on a mixed methodology, and the data were retrieved by multiple researchers who collected data from "158 classrooms of 67 schools in 26 school districts" (p. 33). The study involved 13 states with a geographical spread (Goodlad & Klein, 1970, p. 33). Goodlad and Klein attested that the research staff "made every possible effort to get into three groups of schools"–average schools, innovative schools, and schools that serve poor students (p. 33). Goodlad and Klein described how closely their compiled data matched among multiple researchers with the following

statements:

> Our travelers collected an enormous amount of material, material which provided not only amazing consistency in the quantifiable data but also repetitive impressionistic agreement to the point of analyst boredom. It is this high level of agreement among observers, observers who are themselves teachers, that gives our account of what goes on behind school and classroom doors whatever viability it may possess. (p. 35)

Goodlad and Klein (1970) summarized their findings in eight key points, which have been included in Appendix D. A lack in the quantity of observations and evaluation within teacher evaluation systems was noted in the seventh and eighth points, while the effects of this phenomenon were noted in points one through six. The lack in quality of teacher evaluation systems noted by Goodlad and Klein in points one through six ranged from the inability to improve teaching practices and teacher leadership of teachers to a lack of collaboration of school personnel.

Goodlad and Klein (1970) was one of the first consequential authors to tie the *quantity* and *quality* of teacher evaluation systems together using a large-scale study. Recognizing the lack in quality of teaching practices and teacher leadership in concluding points one through six, Goodlad and Klein recommended two steps. Those steps were laid out in concluding points seven and eight, and they centered on fixing the lack in quantity of observations and evaluation within the teacher evaluation systems.

In their concluding points, Goodlad and Klein (1970) called on principals to improve the teaching practices and teacher leadership by instituting more consistent teacher observation and evaluation within their evaluation systems in order to hone their skills. Goodlad and Klein specifically called on administrators to leave their offices and

enter the “sacred classroom ground” (p. 98).

Goodlad and Klein (1970) also called on teacher schools of education to consistently observe and evaluate future teachers in order to improve their teaching practices and teacher leadership. Goodlad and Klein argued that “neophyte” (p. 105) teachers were rarely observed and evaluated during their training which resulted in a lack of their improvement before they formally entered the classroom. New teachers were not receiving input from sources necessary to ensure they were ready for the classroom.

Goodlad and Klein (1970) provided evidence that a lack in quantity of teacher observation and evaluation within evaluation systems negatively affected teaching practices and teacher leadership. From Goodlad and Klein’s concluding points one through six, the effects of a lack of quantity in observing and evaluating teachers led to ineffective outcomes where teachers were not using quality teaching practices and teacher leadership.

Furthermore, from Goodlad and Klein’s (1970) concluding points seven and eight, evidence was provided that teachers worked autonomously coinciding with Sarason’s (1996) encapsulation theory. Evidence was provided from Goodlad and Klein that teachers were not observed or evaluated consistently which paralleled Schmoker’s (2006) view of inconsistent teacher observations in our present-day systems. Evidence was also provided from Goodlad and Klein’s study that suggested administrators harbored subpar processes related to teaching practices and teacher leadership which mirrored Elmore’s (2000) present-day buffering theory.

Wise et al. (1984) provided more historical evidence that a lack in quantity of teacher observation and evaluation within evaluation systems resulted in the inability to improve teaching practices and teacher leadership. Wise et al. carried out extensive

multiple case studies in which 32 school districts were initially screened and were geographically spread. Four school districts were chosen for a more thorough examination using a case study approach. The initial screening procedures were based on analyzing and gauging the district's purpose for evaluating their teachers, how the districts evaluated their teachers, and the degree of implementation of their teacher evaluation systems (Wise et al., 1984). The four school districts that were examined were Salt Lake City, Utah; Lake Washington, Washington; Greenwich, Connecticut; and Toledo, Ohio (Wise et al., 1984).

Wise et al. (1984) was significant because of the large sample sizes of interview and survey data collected from district officials, administrators, and teachers. The data pertained specifically to a lack in the quantity of observations and evaluation, a lack of quality measurements in observations and evaluation, and the lack of using outcomes from the teacher evaluation systems to improve teaching practices and teacher leadership roles–all within each particular evaluation system (Wise et al., 1984). The study was qualitatively driven and was documented and detailed extensively. Some of the teacher evaluation systems that Wise et al. analyzed dated back to the 1970s in their use.

Pertaining to the significance of the study, Wise et al. (1984) claimed that their case study involved teacher evaluation systems that reflected the interworking of teacher evaluation systems throughout the country. In Lake Washington, Wise et al. observed that their teacher evaluation system "resembles teacher evaluation throughout the country, using virtually the same checklists, the same assessment categories, and the same requirements for pre- and post-observation conferences," and that "in form and structure, then, the Lake Washington evaluation system deviates little from that mandated by the state and from practices in place around the country" (p. 48).

All four sites in Wise et al. (1984) relayed how a lack in quantity of teacher observations and evaluation within evaluation systems inhibited teaching practices and teacher leadership; and when teachers were observed and evaluated consistently, they improved their performance. Teachers were dismissed when they failed to improve their teaching practices and teacher leadership after being consistently observed and evaluated, but Wise et al. detailed it was rare that teachers were consistently observed and evaluated.

Participants in focus groups and interviews warned that there were still incompetent teachers within their schools using ineffective teaching practices and teacher leadership roles (Wise et al., 1984). In all four sites, participants in focus groups and interviews communicated that there was a lack in the quantity of teacher observation and evaluation, and as a result there was not "a real picture of what was going on" (Wise et al., 1984, p. 63). In many cases throughout Wise et al. (1984), teachers relayed that they were not being observed or evaluated consistently within their particular evaluation systems and that this inconsistency in teacher observation and evaluation caused haphazard teaching practices and teacher leadership. For example, one participant in a focus group in Lake Washington noted,

> There is a teacher in this school who only teaches two lessons a year—on the days he is being evaluated. Normally, he does nothing besides drink coffee and read the paper. I resent the fact that bums like him get the same rating I do. There is no room for excellence and it is hard to nail incompetence. (Wise et al., 1984, p. 63)

Wise et al. (1984) also expressed concern that a lack of quantity in observing and evaluating teachers negatively affected competent teachers in that they missed the opportunity to improve their teaching practices and teacher leadership roles. Wise et al. noted this concern by expressing the following:

> Indeed, in all of these systems, principals admit that they spend little time evaluating teachers who appear to be competent; teachers not subject to special treatment [remediation] allege that their evaluations have not given them constructive criticism relevant to their area of teaching expertise. Competent teachers do not necessarily consider the process useless. Rather, they criticize evaluations for providing too few observations and evaluators for making comments that fail to relate specifically to the pedagogical demands of their particular teaching assignments. (p. 52)

The teacher evaluation systems at all four sites in Wise et al. (1984) were failing to heed Darling-Hammond et al.'s (1983) advice in overcoming an "insufficient sampling of performance" (p. 306). McGreal (1983) stressed that the "reliability and usefulness of classroom observation is directly related to the amount . . . of information the supervisor obtains" (p. 98). It was difficult for Wise et al. to rationalize how data from two classroom observations and evaluative judgments based on two samplings of classroom performance could have produced valid and reliable outcomes that would have positive effects on teaching practices and teacher leadership roles (Darling-Hammond et al., 1983).

Marzano, Waters, and McNulty (2005) performed a meta-analysis of 69 principal leadership studies spanning the years 1978 to 2001; involving 2,802 schools; and representing numerous students, teachers, and principals. In describing the leadership behavior of principals and the overall achievement of students Marzano et al. calculated that the estimated average correlation of the 69 studies was 0.25. Marzano et al. also found that one of the many effective leadership behaviors of principals was "continually monitoring the effectiveness of the school's curricular, instructional, and assessment

practices" (p. 56). Based on Marzano et al.'s findings, by ignoring the quantity of observations and evaluations of teachers and not obtaining an accurate picture of what was going on in classrooms evaluation systems, teaching practices and teacher leadership were negatively affected all four sites in Wise et al. (1984) because observers were not capturing necessary data about teacher performance. When teachers have not been observed and evaluated consistently in the past, teaching practices and teacher leadership suffered and students failed to achieve (Marzano et al., 2005; Schmoker, 2006).

McLaughlin and Pfeifer (1988) carried out a case study that paralleled the methodology of Wise et al. (1984) and concluded similarly in terms of the effects of the quantity of observing and evaluating teachers on teaching practices and teacher leadership. The researchers collected data involving four teacher evaluation systems that "were chosen for their commitment to installing meaningful teacher evaluation programs based on accountability and improvement objectives" (McLaughlin & Pfeifer, 1988, p. 89). The researchers spent weeks investigating each teacher evaluation system in the Santa Clara Unified School District, the Mountain View-Los Altos Union High School District, the Moraga School District, and the Charlotte-Mecklenburg School System. The researchers recorded robust qualitative data from interviewing and observing teachers, administrators, and district officials (McLaughlin & Pfeifer, 1988).

At all four sites, McLaughlin and Pfeifer (1988) recorded positive impacts of consistent teacher observation and evaluation on teaching practices and teacher leadership within each particular evaluation system compared to that of Wise et al. (1984), and McLaughlin and Pfeifer noted that teachers were observed in their classrooms more consistently than in the Wise et al. study. McLaughlin and Pfeifer specifically pointed out that the principals in all four sites recognized that consistently

observing and evaluating relayed to teachers that quality teaching practices and teacher leadership were the top priorities. One veteran teacher in the Charlotte-Mecklenburg system summed up this mentality by mentioning,

> How the principal spends his time sends a powerful message to teachers about the priority that something has in the school. The principal serves as a symbol. If he arranges his schedule to spend time on evaluation, then teachers get the message. (McLaughlin & Pfeifer, 1988, p. 37)

The teacher evaluation systems in the McLaughlin and Pfeifer (1988) study were designed for the time-consuming process of observing and evaluating a teacher, and the authors noted that spending more time in the observation and evaluation process improved teaching practices and teacher leadership and provided better student outcomes. This approach paralleled Marzano et al.'s (2005) findings. The principals in McLaughlin and Pfeifer viewed the time-consuming process was worthwhile because "validity and reliability increase because each piece of information sheds additional and substantively different light on goal attainment by the teacher" (p. 42). Observing and evaluating teachers more consistently within each particular evaluation system was found favorable by teachers in all four sites within McLaughlin and Pfeifer, as one teacher expressed,

> The principal has made evaluation an ongoing process this year. . . . There is no territoriality–that is, I don't see my room as mine alone. He is aware, he knows, he sees, and this leads to fairness and reliability. I value his feedback. (p. 130)

Teachers in McLaughlin and Pfeifer (1988) responded favorably to feedback within evaluation systems and improved their teaching practices and teacher leadership. McLaughlin and Pfeifer attributed this phenomenon to a more dedicated approach to consistently observing and evaluating teachers within each particular evaluation system.

Specifically, principals in McLaughlin and Pfeifer admitted to spending more time observing and evaluating teachers than the principals in Wise et al. (1984).

Wise et al. (1984) estimated that principals spent an average of 9% of their time observing and evaluating teachers, and this approach may explain why teachers responded negatively to teacher observation and evaluation and did not improve their teaching practices and teacher leadership roles as a result. In McLaughlin and Pfeifer (1988), principals admittedly spent more time in the classroom observing and evaluating teachers, with some principals in the Charlotte-Mecklenburg site spending 30% of their time observing and evaluating teachers.

Participants in McLaughlin and Pfeifer (1988) reported more positive responses to the effects of teacher evaluation systems on their teaching practices and teacher leadership than did the participants in Wise et al. (1984). Frase (1992) expressed the results of "numerous research studies" that revealed "principals spend only 2.5-10 percent of each work day in classrooms" (p. 178) which contrasted the Goodlad and Klein (1970), Griffith (1973), and Good and Duckett (1980) studies–all of which agreed that the amount of administrative time observing and evaluating teachers was less than 10% of administrative time. Frase and Streshly (1994) advised that principals spend between "40-50% of their day in classrooms . . . working on curriculum and instructional problems" (p. 54) after finding grossly inflated results within teacher evaluation systems.

Furthermore, Frase, Downey, and Canciamilla (1999) argued that it was essential that principals spend 40-80% of their day in the classroom. Frase et al. recommended that principals "spend more time in classrooms, looking and listening for improved ways to do things–wandering with purpose" (p. 36). Frase et al. referred to this approach as Management by Walking Around (MBWA), an approach used in various employee

evaluation models in the business sphere. Frase et al. defined principals who follow the MBWA approach as "MBWA-principals" and identified the top four behaviors these principals share in common:

- conduct classroom walkthroughs;
- observe and work in classrooms;
- participate with teachers in discussions and problem-solving regarding curriculum and instruction; and
- give constructive feedback to teachers regarding curriculum alignment and instructional practices. (p. 37)

Furthermore, Frase et al. (1999) provided research showing how seven desirable outcomes in the educational sphere aligned with principals that follow the MBWA approach:

> Over the past 18 years, a substantial base of research has accumulated linking these MBWA-principal behaviors to highly desired outcomes. Seven desired outcomes have been shown to be closely related to the top-four MBWA activities:
>
> 1. Higher student achievement across socio-economic and cultural lines (Andrews, Soder, & Jacoby, 1986; Andrews, R. & Soder, R. 1987; Heck, 1991; Heck, 1992; Louis and Miles, 1991; Hallinger & Heck, 1996).
> 2. Improved classroom instruction (Teddlie, Kirby & Springfield, 1989).
> 3. Improved teacher perception of principal effectiveness (Andrews & Soder, 1987; Blaze, 1991; Heck, Larsen, & Marcoulides, 1990; Sagor, 1992; Valentine, Clark, Nickerson & Keefe, 1981; Wimpelberg, Teddlie, & Stringfield, 1989).

4. Improved student discipline and student acceptance of advice and criticism (Blaze, 1987; Blaze 1991).
5. Increased teacher efficacy (self-, of other teachers, and of the school) (Chester & Beaudin, 1996; Frase, 1998).
6. Enhanced teacher satisfaction and higher frequency of "flow" experiences (Frase, 1998).
7. Improved teacher attitudes toward teacher appraisal (Frase, 1998). (p. 38)

Turner (1987) concurred with Frase et al. (1999) and reported the findings of a study involving over 1,000 teachers designed to investigate the effects of various teacher evaluation systems on teaching practices and teacher leadership. The teachers completed a survey and participated in interviews, and Turner then compiled the results into a composite narrative that described the overall effects of teacher evaluation systems on teaching practices and teacher leadership based on the responses from teachers.

Based on teacher responses to the survey, the composite of an effective observation and evaluation of teachers involved evaluation systems designed to allow administrators to be a “common sight in the classroom,” to make “many formal and informal visits throughout the year,” to “spend plenty of time observing,” to “know the classroom and students well,” and to be “on hand to point out the teacher’s strengths and weaknesses” (Turner, 1987, p. 40). As a result of an increased quantity of observing and evaluating teachers, “forty-nine percent of respondents said their teaching improved, and twelve percent said it improved significantly, as a result of their evaluations” (Turner, 1987, p. 40).

Turner (1987) also contrasted teachers who perceived an improvement in their teaching practices to those teachers who felt their evaluations either had a negative effect

or no effect on their teaching at all. These teachers composed just over 50% of the sample and described being observed infrequently, such as for 15 minutes once a year if at all. The observations and subsequent evaluations of these teachers were perceived as invalid because they were only based on one short observation or none at all. One teacher responded that

> My last evaluation was very complimentary, but it was based on nothing more than casual observations of how I interact with kids. What if I wasn't an effective teacher? My principal doesn't make in-class evaluations and isn't familiar with the classroom. He doesn't hold conferences; he just writes everyone a complimentary evaluation. For the last two years, he's missed the district deadline and predated his evaluations. (Turner, 1987, p. 41)

Donaldson and Peske (2010) carried out a study similar in approach to Turner (1987). Donaldson and Peske collected data from teachers to identify their perceptions of the effects of various teacher evaluation systems on teaching practices and teacher leadership. Donaldson and Peske's sample included schools from three major charter school operators; and the authors collected qualitative data through interviewing teachers, administrators, and charter school officials. The study sought to uncover the advantages of observing and evaluating teachers in charter schools where evaluation systems were not burdened by a "bureaucratic superstructure" and other constraints, including collective bargaining and teacher tenure (Donaldson & Peske, 2010, p. 6).

Findings in Donaldson and Peske (2010) mirrored Turner's (1987). Specifically, the study highlighted that an increased quantity of observations and evaluation within the evaluation systems of charter schools improved teaching practices and teacher leadership (Donaldson & Peske, 2010). The veteran teachers attested that the evaluation systems in

their prior schools (traditional and charter schools) where teacher observation and evaluation rarely occurred had either no effect or a negative effect on their teaching practices and teacher leadership roles (Donaldson & Peske, 2010).

In contrast, Donaldson and Peske (2010) noted that these particular charter school operators concentrated on "frequent observations" and prioritized "in-the-moment and sustained feedback" (p. 11). As a result, veteran teachers communicated that evaluation systems within these particular charter schools improved their teaching practices and teacher leadership, the latter denoted by Donaldson and Peske as "teacher's non-instructional contributions" (p. 15). The authors made the following observation:

> Overall, the evaluation process that occurs in these three charter school networks differs in several notable ways from that which, according to prior research, occurs in many conventional public schools. Most obvious is the increased frequency of evaluators' informal observations of and feedback on teachers' instruction. As one teacher in the Northern charter school network noted, "We constantly receive feedback." Another teacher said of her school's principal, "You can't get her out of your classroom!"
>
> This feedback is the result of weekly or biweekly observations on the part of evaluators in both the Northern and National charter management organizations. The Western CMO evaluators do not provide teachers with as much observation and feedback, but the organization has created structures to facilitate observation by department heads who visit colleagues' classrooms and provide nonevaluative feedback to every teacher in their department twice per month. (Donaldson & Peske, 2010, p. 28)

The teacher evaluation systems in Donaldson and Peske's (2010) study were

designed to make principals visible entities within classrooms, and teachers improved their teaching practices and leadership roles. Donaldson and Peske explained the successes centered on frequent, informal, and formative classroom observations and evaluation. The teacher evaluation systems Donaldson and Peske were designed to improve teaching practices and teacher leadership rather than use broad, summative judgments based on a lack of data.

Natriello (1983) used a meta-analysis of six studies to investigate the effects of an increased quantity of teacher observation and evaluation within various evaluation systems on the level of teacher internalization. Natriello sought to uncover the extent to which teachers were willing to accept feedback and whether the teachers would change their teaching practices and leadership roles based on the level of frequency of observation and evaluation feedback from administrators.

Natriello's (1983) study was significant because it made conclusions based on data taken from six studies that involved a variety of data collection methods including survey, interview, and comparative studies. The collective studies sampled 58 schools and nearly 800 teachers.

An emergent theme in Natriello (1983) was that teachers in all six evaluation systems were not observed or evaluated consistently with some beginning teachers and veteran teachers never observed or evaluated. Although the studies used various data collection methods, the studies indicated that even when teachers were observed or evaluated, it was either a seldom or occasional event (Natriello, 1983). In the absence of observation and evaluation, teachers expressed freedom from accountability measures; however, Natriello found that the freedom came with a price–teacher loneliness and a bewilderment about whether teachers were using the best teaching practices and forms of

leadership.

Natriello's (1983) analysis involved a statistical calculation of gamma by combining data across the multiple studies of the meta-analysis in order to detect the effects of the frequency of teacher observation and evaluation using the various evaluation systems as an independent variable. Natriello also considered the effects of the independent variables on six dependent variables which included the "teacher perceptions that the evaluations are useful in helping them to improve their performance" (p. 36).

Natriello (1983) concluded overwhelmingly (86 of 100 cases) that the higher the frequency of observation and evaluation within each particular evaluation system, the higher incidence of teacher acceptance of the evaluation and the more likely teachers were to improve their teaching practices and teacher leadership as a result. When Natriello disregarded data from two of the six studies involving smaller sample sizes, there was a 100% incidence of positive outcomes among cases involving an increased amount of observation and evaluation and the teacher's acceptance of the evaluation signifying their willingness to improve.

Administrator and teacher time demands. If an increase in the quantity of observation and evaluation of teachers has historically improved teaching practices and teacher quality, then why have the processes within teacher evaluation systems not met the demand? The answer has involved a shortage of time for administrators to observe and evaluate teachers and a shortage of time for teachers to be involved in the process (Donaldson & Peske, 2010; Goodlad & Klein, 1970; Turner, 1987; Wise et al., 1984). This problem has not only been uncovered in administrative interviews (McLaughlin & Pfeifer, 1988; Wise et al., 1984), but teachers have also acknowledged that administrators have had duties that did not afford them the opportunity to observe and evaluate teachers

frequently enough to improve teaching practices and teacher leadership (Donaldson & Peske, 2010; Wise et al., 1984).

Donaldson and Peske (2010) found that the charter school operators purposely minimized the observer-to-teacher ratio in assigning observation and evaluation duties in order to provide administrators more time for frequent observations and evaluation. It was recognized that "principals are constrained by workloads that thwart their best intentions to provide substantive feedback on instruction" (Donaldson & Peske, 2010, p. 6). As a result, after reducing the observer-to-teacher ratio, teachers admitted that the observation and evaluation process occurred more frequently than in their previous schools and led to improvement in performance as a result (Donaldson & Peske, 2010).

In 2010, Tennessee implemented a new teacher evaluation system as part of winning the RttT competition (TNDOE, 2014). The new teacher evaluation system (Tennessee Educator Acceleration Model [TEAM]) in Tennessee was similar to North Carolina's in terms of structure (both states used Charlotte Danielson's Framework for Teaching observation rubric and both incorporated VAM outcomes provided by the SAS Institute), but TEAM required four classroom observations for veteran teachers and six classroom observations for newer teachers (Reform Support Network, n.d.). This requirement of multiple classroom observations and teacher evaluation in Tennessee was not shared in North Carolina, at least not to the same extent. Veteran teachers in North Carolina have experienced an abbreviated observation and evaluation schedule where, rather than receive a full schedule of observations and evaluation (as in Tennessee), veteran teachers have received two informal classroom observations and were only rated on Standards 1, 4, and 6 (McREL et al., 2012). When veteran teachers in North Carolina have renewed their teaching licenses every 5 years, teachers have received a complete

observation and evaluation schedule.

Although TNDOE designed TEAM with multiple classroom observations in mind, administrators provided feedback relating to the amount of time necessary for multiple classroom observations and evaluation, and some of that feedback indicated that the observation and evaluation demands were too time consuming (Bryant, 2013). Specifically, one principal responded in the *New York Times* that the increase in teacher observation required him to evaluate his stronger teachers as much as his weaker teachers, and that the paperwork involved in the process was a "terrible waste of time" (Winerip, 2011). The principal claimed that the process did not allow him to observe what was going on in his school on a regular basis (Winerip, 2011).

If principals and teachers were consumed with observation and evaluation requirements that were too burdensome, including an exorbitant amount of paperwork or time-consuming preconference and postconference meetings, could principals and teachers have used the evaluation system to improve teaching practices and teacher leadership or was quality shortchanged for quantity? In Tennessee, administrators and teachers provided evidence that an increase in the number of classroom observations and teacher evaluation required both parties to sacrifice valuable time and resources to the process, and administrators were concerned that the time they devoted to classroom observation and teacher evaluation curtailed their ability to provide quality feedback (Bryant, 2013).

Bryant (2013) carried out a mixed-methodology research project in Tennessee studying the perceptions of principals using TEAM and how the implementation of TEAM affected principals and teachers. Participants in Bryant's study included principals across 12 school districts in Tennessee. The principals completed a survey consisting of

26 questions, some of which allowed for qualitative feedback. Bryant noted the following responses (from various respondents) to a question on the survey allowing qualitative feedback:

- There are many obstacles as a building level principal. First, there is not nearly enough time in a day, week, month, or year to get everything done. The TEAM process takes 90% of the instructional day, and principals are forced to stay at their schools until 8:00 or 9:00 at night to get their regular job done.
- Many important things in running a school have to be neglected during the day in order to accommodate the TEAM evaluation system. This is not good for the school or morale.
- It is impossible to implement TEAM fully and maintain the day-to-day operations that are essential to maintain a positive environment that will enhance instruction.
- Time! Time! Time!. . . . It is unrealistic to expect principals to get all this completed and do the daily duties to have a successful staff and student body.
- The biggest obstacle is . . . while it is possible to complete the observations in the expected time frame and still complete all the other responsibilities assigned to school leaders, it provides for us "tunnel vision" of one teacher at a time. While teachers should be able to work on improving, I never really get back to see what changes are made because I am off to watch another teacher. It generally takes me a half of a day to do two observations and that does not include write-ups and the amount of time I have to ponder over the evidence and make a decision based on the rubric. The rest of my day and often well

into the evening is devoted to all the things required of me.

- Time! I cannot prepare evaluation scores and postconference conversations during the school day during my "normal" hours. It all has to be done at night or on weekends. I resent the extra hours I must work to perform the evaluations with fidelity. (p. 102)

Bryant (2013) cited Senge, Kleiner, Roberts, and Roth (1999) and recognized that an inflexible amount of time observing and evaluating teachers may have constrained the ability of administrators to carry out the process with quality and reflect upon how well the system was working. Bryant also referred to the work of DuFour and Marzano (2009) and argued that administrators might have missed the opportunity to collaborate with teachers on a broader scale working "collaboratively with teams to examine evidence of student learning and strategies for improving on those results" (p. 109). If administrators have not had the time to carry out classroom observations and teacher evaluation, they might have been forced to either sacrifice the amount of times they observe and evaluate teachers, or they might have sacrificed quality for quantity in order to meet their quota of classroom observations and teacher evaluation.

Teachers in Tennessee also expressed the lack of time caused by the paperwork, time involved in preconference and postconference meetings, and the amount of time involved in planning for announced classroom observations (Roberts, 2011; TNDOE, 2012). Representative Jim Coley, who taught history at Bolton High School in Shelby, Tennessee, attested to spending 9 hours preparing for an announced classroom observation and subsequent evaluation, and that the paperwork was overwhelming (Roberts, 2011). If teachers were using an exorbitant amount of time to fulfill the requirements for their evaluation and completing the paperwork involved, could they

consistently plan and carry out consistently successful instructional practices for their students? Representative Ron Lollar of Bartlett, Tennessee wanted lawmakers to simplify the teacher evaluation process because when legislation mandating the implementation of TEAM in 2010 passed, teachers were assured that the evaluation process would not take away from teacher's time with burdensome demands (Roberts, 2011). Representative Lollar was concerned about the time it would take for teachers and administrators to complete the evaluation activities within TEAM (Roberts, 2011).

Furthermore, administrators in the pilot data wanted to help guide teachers in improving teaching practices and teacher leadership within the grounded theory analysis of the pilot data for this study involving teacher focus groups and administrative interviews; however, teachers and administrators communicated that there was not enough time. The topic of time dominated the administrative interviews. One principal specifically shared how many teachers each of the four administrators in the school were responsible for in observing and evaluating teaching practices and teacher leadership, and this principal had calculated the amount of time it took to complete the classroom observation and evaluation process to relay the difficulty in scheduling.

This particular principal was concerned about the time commitment involved and whether it was possible logistically to give teacher observations and evaluation the necessary time. Although North Carolina has not required as many classroom observations as Tennessee, both administrators interviewed expressed issues with the amount of time necessary to complete the observation and evaluation of teachers and how the shortage of time may affect the quality of feedback they provided teachers.

Specifically, since the 2011-2012 school year, North Carolina teachers have been required to receive yearly evaluations under NCEES, but career status teachers have been

observed and evaluated based on an abbreviated form and received full evaluations upon renewal of their teaching licenses (every 5 years) (Batten, Britt, DeNeal, & Hales, 2012). This was a policy change from an earlier approach where North Carolina teachers only received full evaluations once every 5 years upon renewal of their teaching licenses (Batten et al., 2012). Beginning teachers in North Carolina receive full, formal evaluations under NCEES (Batten et al., 2012).

Batten et al. (2012) collected data across North Carolina and analyzed data involving statewide teacher evaluation scores from NCEES for approximately 46,000 teachers with a final dataset of 11,430 teachers. Batten et al. was significant to this particular study for various reasons (which will be shown later), and the results were interspersed throughout this current study. Batten et al. mentioned this important note transmitting time issues surrounding the processes involved in NCEES:

> Finally, we believe it is important to discuss the role of administrators in the performance evaluation process. With recent budget constraints, both principals and assistant principals are taking on more duties. Proper evaluations are necessary for accountability, but require administrators to take an appropriate amount of time for each evaluation. Yet administrators' time is becoming more and more constrained. This could become a critical issue as policy shifts for administrators to evaluate all teachers during each school year. Administrators must view performance evaluations as a priority and be given both the time and the support to devote to this task. (p. 8)

In the pilot data, there was some consistency among teachers in focus groups perceiving an increase in the number of classroom observations and evaluation after the implementation of NCEES. Although there was no specific question designed for the

focus group teachers on this topic, teachers commented that they perceived an increase in receiving classroom observations and evaluation after NCEES was implemented. Teachers in the focus group agreed that historically there was a lack of quantity in their observations and evaluations (some said they were rarely observed during instructional time). No teachers in any of the focus groups mentioned they experienced less classroom observation and evaluation after NCEES was implemented.

The pilot data also revealed that teachers were frustrated with the time some postconferences required, but this sentiment was not consistent among all teachers. Some revealed it depended upon the administrator assigned to them. One teacher specifically stated that her postconference took 2 hours because the administrator required her to show artifacts to prove she met certain standards on which she was initially rated low. The teacher expressed a concern that the postconference was not about improving her performance, rather it was a "waste of time trying to prove to him that I really did what the rubric wanted me to do." This same teacher said she felt as though her evaluation was not about what she did right or wrong during the lesson or could have done better, but the evaluation centered on whether she met the criteria on the evaluation rubric which proved to be a time-consuming process.

The first category, "Quantity of Classroom Observations and Teacher Evaluation" (as depicted in Figure 2), was inferred initially from pilot data supplied by administrative interviews and from teachers participating in focus groups using a grounded theory approach. Further investigation ensued spanning literature from multiple decades, and that data combined with data from the grounded theory analysis were used to construct the categories depicted in Figure 2.

The results revealed that the amount of time administrators spent in classrooms

mattered, specifically that the more time administrators spent in the classrooms observing and evaluating teachers, the more teachers improved upon their teaching practices and teacher leadership. It was also found that time constraints have negatively influenced the ability of teachers and administrators to fully participate in using the processes within teacher evaluation systems to improve teaching practices and teacher leadership.

The Effects of the Quality of Classroom Observations and Teacher Evaluation

Validity of teacher evaluation systems. The quantity of classroom observations and teacher evaluation has shown to be inextricably linked with the quality of the classroom observations and teacher evaluation (McGreal, 1983). In order to have successful observation and evaluation of each teacher in a school, McGreal (1983) advised that principals spend at least "15 to 25 hours in a classroom over several months" and be able to "understand the experiences that pupils and teachers were having, and not simply describe or count the behaviors they were displaying" (p. 32). Without quality classroom observations and teacher evaluation, the amount of time a principal spends in a teacher's classroom could result in diminished effects when attempting to improve teaching practices and teacher leadership (McGreal, 1983).

McGreal (1983) cited Eisner (1982) arguing that to improve teaching practices and teacher leadership within teacher evaluation systems, principals should use their "educational connoisseurship" (p. 32)–a deep, sensitive, and perceptible understanding of the complexity of the classroom. A teacher evaluation system that concentrated solely on quantity may jeopardize the quality of observing and evaluating teachers, which would negate the ability of a teacher evaluation system to positively affect teacher performance and growth. When observing and evaluating teachers, the quality of the processes involved in teacher evaluation systems has been shown to be as important as the quantity

in improving teaching practices and teacher leadership (Peterson, 2000).

Teacher evaluation systems have used data collection instruments in the form of checklists, and McGreal (1983) noted these checklists were used to keep count of minimum competencies teachers needed to perform in order to maintain employment rather than to improve teaching practices and teacher leadership. Early in the history of teacher evaluation systems, Thompson's (1962) "intangibles" were widely part of data collection instruments. Teachers were judged based on such things as their appearance, voice, warmth, and enthusiasm without concentrating on how effective teachers taught within the classrooms or led within their schools (Danielson & McGreal, 2000).

Peterson (2000) cited Wood and Pohland (1979) and reported that in analyzing 65 rating scales from teacher evaluation systems at that time, 28% of the items directly dealt with teaching practices and teacher leadership (e.g., "adequate preparation of lesson plans"; "techniques in evaluation of student progress"; "challenges students to think, inquire, and analyze"), 30% "focused on personal characteristics," while 1% focused on student outcomes (p. 22). The revamping of current teacher evaluation systems since the implementation of the RttT initiative has concentrated more on teacher performance and student outcomes (Danielson, 2007; Hanover Research, 2011).

Although collecting data on superficial characteristics of teachers may give some baseline for judgment within teacher evaluation systems, historically the checklists were never shown to measure and improve teacher performance or student achievement (Danielson & McGreal, 2000; Peterson, 2000). Peterson (2000) cited Cook and Richards (1972) who found that the checklists were "more of a reflection of the rater's point of view than of a teacher's actual classroom behavior" (p. 16). There was little historical evidence that teacher evaluation systems used quality data collection instruments that

positively affected teaching practices and leadership (Peterson, 2000).

Recently designed teacher evaluation systems have been shown to improve teaching practices and teacher leadership and increase student learning because they have been built on quality classroom observations and multiple teacher evaluation ratings (Danielson, 2007; Marzano, Toth, & Schooling, n.d.; Stronge & Associates, 2013). The positive outcomes of newly created teacher evaluation systems under the RttT initiative have already been noted in analyzing the quantity of classroom observations and teacher evaluation in the last section of this study and will continue to be included interspersed within this section while exploring the quality of teacher observation and evaluation within historical and current contexts.

Historically, Soar et al. (1983) called for reform in teacher evaluation systems after finding that multiple studies did not provide any evidence that the observation and evaluation ratings of teachers coincided with quality measurement, the improvement of teachers, or an increase in student achievement. Soar et al. remarked that "each study reported the same conclusion: the ratings of teachers' performance made by reasonably sophisticated observers had no validity as predictors of teacher effectiveness" (p. 244).

Furthermore, Coker, Medley, and Soar (1980) carried out a study in which 60 teachers were observed and rated using a classroom observation instrument designed from other longstanding classroom observation instruments. The classroom observation instruments used to design Coker et al.'s instrument were in use throughout various states for the previous 20 years and considered valid and reliable. The teachers in Coker et al. were observed for 2 years, and the results were compared to student achievement gains. Pearson product-moment correlations were calculated for each component on the classroom observation instrument compared to the pupil gain measures, and the level of

significance of each correlation was calculated separately for each year using a chi-squared table (Coker et al., 1980).

Coker et al. (1980) found that most of the correlations were nonsignificant; and of the 13 that were significant (at a 5% probability level), five were negatively correlated with one significant relationship being positive in the high school grades but negative in the elementary grades. Coker et al. concluded that the results were "startling in the mixed and negative support they offer for our best ideas about how an effective teacher of cognitive skills and content behaves (p. 149).

When student achievement has not aligned with outcomes of rating instruments used within teacher evaluation systems, the quality of the system has been shown to be at stake. Coker et al. (1980) complained that teacher evaluation systems could be doing as much harm as good in improving student achievement because there was no validity in determining how, when, or why principals observed and evaluated teachers.

In 2009, TNTP identified major problems with the validity of classroom observation and teacher evaluation with teacher evaluation systems across the country (Weisberg et al., 2009). Weisberg et al. (2009) found that teacher ratings from teacher observations and evaluation were grossly inflated. In a binary teacher evaluation system (employing two ratings such as "satisfactory" or "unsatisfactory"), Weisberg et al. found that 99% of teachers in the systems analyzed were rated as "satisfactory" over a span of years. Weisberg also found that in a system using multiple ratings, 94% of teachers received the highest ratings over a span of years.

The outcomes were found to be skewed, and this was especially so when considering the low academic achievement of the students noted by Weisberg et al. (2009). Among four states and 12 school districts involved in the study, Weisberg et al.

cited Denver public schools where 98% of teachers were rated "satisfactory," and the district failed to make Annual Yearly Progress under the stipulations of the NCLB Act of 2002.

Weisberg et al. (2009) was significant because it involved 15,000 teachers; 1,300 administrators; and 80 local and state officials. Weisberg et al. showed the lack of quality in teacher evaluation systems from the smallest districts to the largest. Weisberg et al. advised engaging in valid teacher observation and evaluation practices where excellence in teaching was recognized for the sake of promotion, training, and compensation.

While Weisberg et al. (2009) detailed the issues surrounding observation and evaluation inflation, newly devised teacher evaluation systems have been designed to identify teachers based on their effectiveness in increasing student achievement based on the performance of teachers. An important basis of newly designed teacher evaluation systems is the data collection instruments and rubrics that have been designed based on in-depth research involving the correlation of teaching practices and teacher leadership and their relationship to student outcomes (Danielson, 2007). This approach has been shown to avoid teacher inflation ratings and differentiate among teachers based on their performance (Daley & Kim, 2010). The effects have been noted in the literature and have harvested positive results in teachers being differentiated based on their performance.

Differentiation of teachers based on performance. For instance, Daley and Kim (2010) studied the effects of the teacher evaluation system TAP (The System of Teacher and Student Advancement) on teaching practices and teacher leadership. The TAP system was already in place before the RttT initiative of 2009, but this teacher evaluation system shared the same focus on rating teachers based on their performance and the outcomes of student achievement. TAP served 7,500 teachers at the time of Daley and Kim's study,

and the aggregate average teacher ratings were significantly different than that of Weisberg et al. (2009).

On a scale of 1 to 5, the average rating was 3.5 in data from the Daley and Kim (2010) study, and the shape of the data graphed was bell-shaped rather than skewed toward higher ratings. This meant the vast majority of teachers were average performers which was an expected statistical distribution for teacher performance (Daley & Kim, 2010).

Daley and Kim (2010) was significant because it showed the quantitative distribution of a large sample of teachers based on their performance, and the study showed how the distribution of teachers based on performance differed from historical distributions. Specifically, Daley and Kim's results in Appendix E were significantly different compared to Weisberg et al.'s (2009) found in Appendix F. Weisberg et al. revealed the historical inflation of teacher ratings while Daley and Kim showed TAP's ability to differentiate among teachers, which was important for principals to offer specific feedback to teachers rated at lower ends of the spectrum. Without such a distribution, it has been difficult for teacher evaluation systems to determine which teachers needed help (Peterson, 2000).

This same phenomenon was observed after the District of Columbia implemented a new teacher evaluation system (IMPACT) during the 2009-2010 school year. Curtis (2011) noted the validity of the aggregate average teacher ratings after IMPACT was implemented where instead of 95% of teachers being found "effective" or better (as was the case under the previous teacher evaluation system), 81% of teachers were found to be "effective" or better. Curtis (2011) found a significant shift from the number of teachers being found less than effective, congruent to "meets expectations," from 5% to 19%. A

full breakdown of the results for the Curtis (2011) study is found in Appendix G. The results of Curtis paralleled Daley and Kim (2010).

Alongside the validation of IMPACT in teacher distribution, Curtis (2011) identified the accountability and rewards of IMPACT apparent for teachers scoring less than "effective" and also on teachers scoring "highly effective." Teachers scoring "ineffective" were in danger of being dismissed within 1 year of the rating without improvement, and teachers scoring "minimally effective" had 2 years to improve their rating to "effective" or be in danger of dismissal (Curtis, 2011). Teachers scoring "highly effective" were eligible for higher compensation based on their performance.

The differentiation of teacher performance using classroom observation and teacher evaluation rubrics has been crucial to validating the process within teacher evaluation systems (Curtis, 2011; Daley & Kim, 2010; Weisberg et al., 2009). Without being able to use valid processes to identify the top-performing teachers, it was difficult to learn from them in order to improve teaching practices and teacher leadership (Weisberg et al., 2009).

Likewise, without being able to use valid processes to identify the middle and low performers, it was difficult to offer them professional development opportunities in order to improve teaching practices and teacher leadership (Weisberg et al., 2009). Curtis (2011) and Daley and Kim (2010) showed how recently designed teacher evaluation systems have implemented performance-based standards to provide stakeholders with valid data identifying teachers to either professionally develop or offer rewards. This approach has been a catalyst in improving teaching practices and teacher leadership and learning from high performers (Curtis, 2011; Daley & Kim, 2010; MET, 2010; MET 2012; Sartain et al., 2011).

Batten et al. (2012) and the Consortium for Educational Research and Evaluation–North Carolina (CERENC) team (Lynn, Barrett, Marks, Comperatore, Henry, & Gurthrie, 2013) provided a similar analysis into the distribution of teacher ratings in North Carolina compared to that of Weisberg et al. (2009), Daley and Kim (2010), and Curtis (2011). Batten et al. calculated the average North Carolina teacher evaluation outcome for N=11,430 teachers from the 2010-2011 school year stemming from NCEES. Batten et al. (2012) was significant because it provided the distribution of teacher ratings and the variation of those ratings in standard deviation units for teachers in North Carolina. Batten et al. collected NCEES teacher ratings for average evaluation scores statewide, assigned numerical values to each NCEES standard, and provided a composite graph listed in Appendix H. Lynn et al.'s (2013) analysis mirrored Batten et al.'s and is included in Appendix I.

Lynn et al. (2013) compiled data similar to Batten et al. (2012) for the 2011-2012 school year for a slightly smaller sample size (N=10,616). While Batten et al. calculated the statewide mean performance score for teachers on Standards 1-5 of NCEES, Lynn et al. showed the distribution of teacher ratings between each of the five standards for teachers in North Carolina.

Batten et al. (2012) showed NCEES teacher ratings provided by administrators were bimodal, indicating administrators had a tendency to rate teachers with four "proficient" ratings and one "accomplished" rating (resulting in an average score of 3.2) or four "accomplished" ratings and one "distinguished" rating (resulting in an average score of 4.2). Batten et al. also observed that this same distribution was similar for each individual standard in NCEES.

Batten et al. (2012) found the statewide mean teacher evaluation score was 3.6

with a standard deviation of 0.6 of one point. This meant the majority of teachers in Batten et al. received ratings from NCEES clustered around the mean.

This outcome has had a negative effect on teaching practices and teacher leadership because the reduced amount of variation made it difficult to offer professional development to teachers who needed it. Within this distribution, it made it difficult to identify those teachers who had shown the highest quality of teaching practices and teacher leadership in order to reward them and learn from them. Even so, Batten et al.'s (2012) distribution using NCEES was an improvement in the aim to improve teaching practices and teacher leadership compared to that of Weisberg et al.'s (2009) distribution where outcomes were heavily skewed towards the high end of the distribution.

Lynn et al. (2013) also analyzed effects of NCEES on teaching practices and teacher leadership for both principals and teachers in North Carolina. Lynn et al. was significant because it measured the distribution of teachers among standards based on principal ratings; it measured the extent Standard 6 EVAAS® outcomes related to Standards 1-5 of NCEES for teachers rated "accomplished" or "distinguished"; and it measured the relationship between Standard 6 EVAAS® outcomes and all principal ratings for teachers ("proficient" → "distinguished") on Standards 1-5 of NCEES. Lynn et al. also investigated the perceptions of teachers surrounding Standard 6 EVAAS® data and how NCEES outcomes were used to improve teaching practices and teacher leadership. The results of Lynn et al. are interspersed throughout this study where applicable.

Data from Lynn et al. (2013) reinforced Batten et al.'s (2012) conclusion that teachers were tightly clustered around the mean resulting in low variant outcomes. As

with Batten et al., the outcomes from Lynn et al. showed an improvement from previous teacher evaluation systems in Weisberg et al.'s (2009) study in that there was a wider distribution of teacher performance at the bottom and top tails rather than having teachers skewed solely at the top tail.

Daley and Kim (2010), Curtis (2011), Batten et al. (2012), and Lynn et al. (2013) noted an improvement in the ability of teacher evaluation systems to differentiate among teachers based on the quality of their teaching practices and teacher leadership, an improvement in the ability to offer professional development to teachers who needed it most, and an improvement in the ability to reward top-performing teachers. However, there has been evidence of the negative effects when the distribution of teacher performance contradicts within teacher evaluation systems. When comparing principal ratings versus student outcomes (using VAM), teacher evaluation systems in Tennessee have shown some negative effects of uncorrelated measures of teacher performance on the ability to measure teaching practices and teacher leadership with quality.

In 2011, Tennessee implemented a "comprehensive, student outcome-based statewide educator evaluation system," which "was a key tenet of Tennessee's *First to the Top Act* (FttT)" legislation adopted by Tennessee's General Assembly in 2010 (TNDOE, 2012, p. 1). The law required that 50% of a teacher's evaluation be based on principal classroom ratings and 50% on student growth measures–35% based on TVAAS® (or similar VAM measure) and another 15% based on growth from a mutually (teacher and principal) chosen growth measure (TNDOE, 2012).

In a mixed-methodology study, the TNDOE (2012) study attempted to validate its teacher evaluation systems by analyzing principal ratings versus TVAAS® outcomes. The results are shown in Appendix J.

The TNDOE (2012) study noted that while the TVAAS® outcomes provided differentiated outcomes in respect to a wide range of teaching practices and teacher leadership, principal ratings using the four state-approved performance rubrics could have been inflated following the trend reflected in Weisberg et al.'s (2009) research. Specifically, the TNDOE (2012) study found that on average, "for a teacher with an individual value-added score of a 5," the average principal rating was just above a 4; whereas "teachers with a value-added score of a 1 received an average observation score of a 3.64" (p. 32).

The TNDOE (2012) study outlined the contradiction in teachers receiving low VAM outcomes and high principal ratings within its teacher evaluation system and the negative effects on teaching practices and teacher leadership that were highlighted by such a situation:

> Less than one half of one percent of teachers are identified by their evaluators as falling significantly below expectations. At the same time student growth data identifies more than 16 percent of such teachers. This creates an environment in which struggling teachers receive little assistance or feedback on how to improve. In many cases, evaluators are telling teachers they exceed expectations in their observation feedback when in fact student outcomes paint a very different picture. This behavior skirts managerial responsibility and ensures that districts fail to align professional development for teachers in a way that focuses on the greatest areas of need. This in turn leads teachers to maintain the same instructional methods and strategies and results in continued low levels of growth for their students. This is unacceptable for students assigned to these teachers, since they will, in all likelihood, fall behind their peers who are assigned to more effective

instructors. (p. 32)

One caveat in using the TNDOE (2012) study within the present context of this subsection was that TNDOE did not include a calculated correlation statistic like the upcoming studies to be featured within this subsection. Without an overall correlation measurement, it has been difficult to conclude in terms of the validity of the principal ratings (using the four rubrics) in measuring student outcomes from VAM in Tennessee.

Although the percent of principal ratings were skewed towards the higher ratings, the question would still remain whether or not the higher correlation levels at the higher rating levels would be enough to offset the lower correlation levels at the lower rating levels where principal ratings and VAM did not align. For this reason, the data provided in Appendix J did not necessarily invalidate Tennessee's teacher evaluation system; it just uncovered differences in overall measurement between two different devices without looking at the specific correlation between each standard.

Also, the percent differences did not explain which measure was invalid, just that they differed. In other words, it has not been possible to determine whether principal ratings or VAM outcomes were off target in Tennessee, but it has been possible to conclude the two measures did not align overall, especially at the low end. The authors of the TNDOE (2012) study tended to lean more towards principal judgments having been off target when they remarked, "this behavior skirts managerial responsibility" (p. 32). It may have been correct for the authors of the TNDOE (2012) study to make that assumption based on data from Weisberg et al. (2009), Daley and Kim (2010), Curtis (2011), and Sartain et al. (2011). The aforementioned authors have shown that administrators in revamped teacher evaluation systems have differentiated among teachers based on performance where much larger proportions than 0.2% of teachers

were found in the lowest categories such as was found in Tennessee (see Appendix J).

While the TNDOE (2012) study showed the danger of misalignment between principal judgments and VAM, validating teacher observation and evaluation rubrics within teacher evaluation systems has become an important exercise for researchers to measure the effects on teaching practices and teacher leadership (Curtis, 2011; Danielson, 2007; Marzano et al., n.d..; Stronge & Associates, 2013). If teachers were to improve teaching practices and teacher leadership, there should be a valid judgment of a teacher's performance within teacher evaluation systems.

Batten et al. (2012) found similar outcomes in North Carolina using NCEES teacher ratings and EVAAS® outcomes compared to the TNDOE (2012) study. The goal of Batten et al.'s study was to measure the relationship of Standard 6 EVAAS® data to the mean Standard 4 score from NCEES and to a composite mean score across all standards of NCEES. Batten et al. concentrated their analysis on Standard 4 of NCEES because it had the highest mean score and it was the standard that rated the quality of teaching practices in the classroom. Batten et al. assumed Standard 4 of NCEES should have directly contributed to student achievement because the standard is designed to measure the quality of teaching practices in the classroom during classroom observations. Batten et al. ran regression models for 35 LEAs, eight North Carolina regions, and at the state level to measure R^2 values signifying the level of effects explained by the model versus variation from random factors.

In analyzing NCEES ratings for over 11,000 teachers, Batten et al. (2012) found the relationship of student growth in Standard 6 of NCEES (from EVAAS®) to both Standard 4 and the mean NCEES teacher ratings was smaller than expected. Batten et al.

found that in the highest of the 35 LEAs, EVAAS® data predicted 9% of the variation in Standard 4 of NCEES (with 91% resulting from unexplained or random factors) and 13% of the variation of the mean teacher evaluation score.

The results for Batten et al. (2012) did not improve at the regional or state levels. The two tables provided in Appendices K and L summarize Batten et al.'s findings at the regional level in respect to EVAAS® outcomes versus Standard 4 ratings and EVAAS® outcomes versus the mean teacher evaluation ratings.

Batten et al. (2012) found an R^2 value of 8.9% between EVAAS scores and the mean teacher rating in NCEES with the remaining "91 percent of the variation resulted from unexplained or random factors" (p. 5), but this outcome was for the LEA with the highest correlation between EVAAS® and the mean teacher rating in NCEES; the lowest LEA had an R^2 value of 3.7% where the remaining 96.3% would have resulted from unexplained or random factors (Batten et al., 2012).

Batten et al.'s (2012) outcome paralleled Lynn et al. (2013) in which the R^2 value was about 6% for the ratings alongside experience, gender, and ethnicity in explaining the variability of EVAAS® outcomes. Lynn et al. provided several possibilities in explaining the low connection among the various independent variables in explaining the variability of teacher EVAAS® outcomes. Lynn et al. hypothesized that the low R^2 values could have stemmed from EVAAS® measuring something other than what principals were basing their ratings on, that there could have been a lack of reliability in principal ratings and/or EVAAS® outcomes, or that principal ratings were clustered tightly around the mean as Batten et al. observed. Batten et al. elaborated more specifically on both the phenomenon observed and the cause:

> We classified teachers based on EVAAS® Teacher Effect scores from highest to lowest and found that, out of more than 11,000 teachers, the average performance evaluation rating of the 100 least effective teachers was a 3.2, while the average evaluation score for the 100 most effective teachers was only a 3.8. In other words, out of more than 11,000 teachers, the 100 whose instruction contributed the most to student growth and the 100 whose instruction contributed the least to student growth were all rated somewhere between "Proficient" and "Accomplished" on their performance evaluation ratings.
>
> We suspect the low degree of variation in evaluation scores may partially explain the weak relationship between student growth data and teacher evaluation ratings. If administrators give similar scores to all their teachers, a high performing teacher would not likely stand out in terms of their performance evaluation. Conversely, if an administrator gives a wide range of scores, he or she would likely be particularly careful to ensure that high scores go to the most effective teachers. (p. 6)

As with the TNDOE (2012) study, Batten et al. (2012) and Lynn et al. (2013) did not necessarily invalidate NCEES–a misalignment may have occurred where there was an insignificant distribution of teacher ratings within NCEES. The misalignment may have caused little variation in teacher ratings and low associative values with EVAAS® data at the lowest rating levels. The possible misalignment within teacher evaluation systems has had negative effects on teaching practices and teacher leadership when the validity of teacher evaluation systems became suspect (Batten et al., 2012).

Both in Tennessee and North Carolina, the TNDOE (2012) study, Batten et al. (2012), and Lynn et al. (2013) showed that there should have been agreement among

components of teacher evaluation systems if there would be fidelity in the process of identifying struggling teachers to improve their teaching practices and teacher leadership. Batten et al. and Lynn et al. advised targeted training for principals enabling them to produce a wider range of ratings, specifically to identify developing teachers.

Construct validity within teacher evaluation systems: linking teacher ratings and student outcomes. Stronge & Associates (2013) specifically identified the validation process like those of the TNDOE (2012) study and Batten et al. (2012) as "construct validity": "construct" as an "attribute, proficiency, ability, or skill, and is defined by established theories"; and "validity" as "the extent to which a measure actually measures the construct it is meant to measure" (p. 8). More specifically, in terms of teacher observation and evaluation rubrics versus student achievement, Sartain et al. (2011) referred to their analysis of the validity of Danielson's (2007) *Framework* as "the relationship between classroom observation ratings and student achievement" (p. 10). Student achievement was measured using TVAAS® outcomes in the TNDOE (2012) study and using EVAAS® outcomes in Batten et al.

Sartain et al. (2011) used a large sample of teachers (N=417 reading teachers with 795 observations, N=340 mathematics teachers with 653 observations) to show the "significant relationship between observation ratings and value-added measures" within Chicago's teacher evaluation system in Chicago Public Schools (CPS) (p. 56). Using data compiled and appearing in Appendix M, Sartain et al. showed there was a significant correlation between the average teacher's rating for components of Danielson's (2007) *Framework* and their value-added outcomes, other than for "2a" (for both reading and mathematics).

For example, Sartain et al. (2011) noted that a teacher who was rated

"unsatisfactory" in Component "3e" in reading had on average -0.397 student growth, which translated into more than one-third less growth than a teacher who had students who achieved average growth (which was set at "0"). A "distinguished" teacher in that same component would produce more than one-third higher growth than an average teacher.

Lynn et al. (2013) provided data for North Carolina teachers that paralleled Sartain et al.'s (2011) findings in CPS. For every movement up in NCEES principal ratings, there was an improvement in EVAAS® outcomes; and Lynn et al. controlled for ethnicity, gender, and average teaching experience. Exceptions existed in the "not demonstrated" category, but the sample sizes for this category were too small to be useful (Lynn et al., 2013). Lynn et al. collected these data and organized them as shown in Appendix N.

There were useful trends uncovered in Lynn et al. (2013) apparent in the data in Appendix N. White teachers were overrepresented toward higher ratings across all standards, while males were underrepresented toward higher ratings across all standards (Lynn et al., 2013). Another important trend was that more experienced teachers tended to be rated high across all standards (Lynn et al., 2013).

Lynn et al. (2013) provided a basis in the primary research methods of this study to collect data simliar to Lynn et al. and investigate the effects of NCEES on teaching practices and teacher leadership based on the demographic groups presented within Lynn et al.'s analysis. The demographic groups included gender, ethnicity/race, and years of teaching experience. The goal was to investigate whether NCEES was fairly providing outcomes across demographic groups to improve teaching practices and teacher leadership.

The results of Sartain et al. (2011) and Lynn et al. (2013) were important because they provided validity in identifying more or less effective teachers using multiple measures within teacher evaluation systems (principal ratings and VAM outcomes). Education researchers have been interested in finding the correlation of principal ratings and student achievement within teacher evaluation systems in order to identify effective teachers and replicate their practices (Lynn et al., 2013; Sartain et al., 2011). It has also been important to researchers to identify less effective teachers and provide professional development.to improve teaching practices and teacher leadership (Sartain et al., 2011). This process establishes validity in principal ratings versus student outcomes and has linked the ability of teacher evaluation systems to improve teaching practices and teacher leadership while increasing student achievement (Sartain et al., 2011).

Early attempts to observe and evaluate teachers using teacher evaluation systems and to investigate the effects of teaching practices and teacher leadership on student achievement proved difficult because of the inability to tease out the correlation of teacher ratings and their effects on student learning (Peterson, 2000). During the 1980s and 1990s, significant improvements were made to teacher observation and evaluation instruments within teacher evaluation systems. There was a transition away from simplistic categorizations such as Griffith's (1973) sample checklist for teachers in New York's teacher evaluation system that consisted of categories such as "very good," "good," "fair," and "unsatisfactory," to a more standards-based movement (Peterson, 2000).

Later categorizations of teacher performance within the newly devised standards-based movement used ratings such as "above standard," "at standard," and "below standard." Other rating schemes concentrated on teacher growth such as those from

Danielson's (2007) *Framework* and included "unsatisfactory," "basic," "proficient," and "distinguished." These rating categories were designed to capture teacher performance within teacher evaluation systems based on agreed standards of performance and were based on evidence stemming from research (Danielson, 2007). This approach improved the quality of instruments designed to measure teaching practices and teacher leadership in the classroom (and outside the classroom) within teacher evaluation systems which translated into high student academic achievement (Danielson, 2007).

Danielson's (2007) *Framework for Teaching* has become popularized nationally as an accepted instrument to gauge the quality of a teacher within teacher evaluation systems. Danielson (2007) asserted that the *Framework* "identifies those aspects of a teacher's responsibilities that have been documented through empirical studies and theoretical research as promoting improved student learning" (p. 1). Danielson's (2007) *Framework* has attracted research attention because it has been the basis for which NCEES has been designed including observation and evaluation instruments and rubrics.

In order to ascertain the quality of Danielson's (2007) *Framework* within teacher evaluation systems in terms of its validity, Milanowski, Kimball, and White (2004) investigated its impact on teaching practices and teacher leadership in three systems located in Los Angeles (Vaughn); Cincinnati, OH; and in Nevada (Washoe). The study was significant in that it analyzed the quality of Danielson's (2007) *Framework* within many large-scale school systems and showed how teacher evaluation systems incorporating the *Framework* affected large groups of teachers in improving teaching practices and teacher leadership.

Specifically, Milanowski et al. (2004) aimed to describe the validity of the Framework by measuring the degree of correlation between VAM outcomes at the

student level and the ratings of their teachers under the *Framework* rubric. There was a previous round of data (Round 1) from a previous year which the authors included in Milanowski et al. (2004). The results of Round 1 and Round 2 of Milanowski et al. (2004) data are displayed in Appendices O and P, respectively, noting that the results were weak to moderately positive other than in Cincinnati (for science) where the confidence interval was included at "0."

Milanowski et al. (2004) also calculated the effects of a more highly rated teacher (based on the *Framework's* rubric) on student achievement using VAM outcomes in an attempt to validate the evaluation system. Milanowski et al. (2004) calculated the number of standard deviations that were observed with a change in a teacher's rating from one level to the next highest level (i.e., from basic to proficient). Milanowski et al.'s (2004) analysis allowed for comparisons across grades, subjects, and sites; and the results are shown in Appendix Q. Milanowski et al. (2004) commented that

> These effects, though small to moderate in size, could add up to a substantial advantage for a student with two or three consecutive teachers rated at the "distinguished" rather than the "proficient" level, or the "proficient" rather than "basic" level. (p. 14)

Milanowski et al. (2004) concluded that the teacher evaluation systems incorporating the *Framework* led to valid outcomes in measuring student outcomes compared to teacher performance because of the correlation of teacher ratings and student outcomes. Making such measurements highlighted the teaching practices and teacher leadership that could be replicated elsewhere and used to professionally develop struggling teachers to improve teaching practices and teacher leadership. Milanowski et al. (2004) also found that although the *Framework* required a "more intensive collection

of evidence" (p. 2), the benefits would give education leaders a more valid basis for teacher licensure renewal. Another advantage of the approach used by Milanowski et al. (2004) would give teacher evaluation systems the ability to differentiate feedback to struggling teachers in order to improve teaching practices and teacher leadership. The level of support for teachers at the lowest ratings could be planned, implemented, and measured in order to improve teaching practices and teacher leadership.

Heneman, Milanowski, Kimball, and Odden (2006) also carried out a study that measured student gains based on VAM outcomes and how they related to a teacher's rating from Danielson's (2007) *Framework*. The authors used VAM outcomes while controlling for student characteristics that affected student learning, such as a student's socioeconomic status (Heneman et al., 2006).

Similar to Milanowski et al. (2004), Heneman et al. (2006) was significant because it analyzed Danielson's (2007) *Framework* within many large-scale school systems. Heneman et al. measured the validity of teacher evaluation systems by showing the effects of the *Framework's* observation and evaluation rubric on large groups of teachers in their efforts to improve student achievement based on the quality of their teaching practices and teacher leadership. Unlike Milanowki et al. (2004), Heneman et al. (2006) combined a 3-year rolling average in order to provide more validity in outcomes. The results of Heneman et al. (2006) are included in Appendix R for consideration within the context of the findings.

Based on the data, Heneman et al. (2006) found positive correlation relationships when comparing the teachers' *Framework* ratings within the teacher evaluation systems of the various sites and their students' VAM outcomes, with some of those outcomes considered substantial–namely in Vaughn and Cincinnati. Heneman et al. (2006)

speculated that Cincinnati and Vaughn had higher correlation relationships because of the attention those sites gave to using multiple evaluators to observe and evaluate teachers (which may have attributed to more valid ratings) rather than single evaluators at the other sites.

Heneman et al. (2006) showed that the evaluation systems using the *Framework* used a quality teacher observation and evaluation rubric in measuring teaching practices and teacher leadership and that the rubrics correlated with student achievement. The results also showed that principals were able to accurately judge teacher performance using the *Framework* (Heneman et al., 2006).

Heneman et al.'s (2006) approach was important because it validated the ability of principals to identify teachers with high-level teaching practices and teacher leadership within quality teacher evaluation systems built on quality rubrics. By identifying higher and lower performing teachers based on a validated teacher evaluation system, school leaders then focused on improving teaching practices and teacher leadership for lower performing teachers (Heneman et al., 2006).

Milanowski et al. (2004), Heneman et al. (2006), and Sartain et al. (2011) supported an inference used within the primary research methods of this study (Chapter 3) measuring the improvement of teachers within the NCEES rating system. Since Milanowski et al. (2004), Heneman et al., and Sartain et al. showed an improvement in student outcomes in conjunction with an improvement in teacher ratings using the rubric from the *Framework*, it could be inferred within the parameters of the primary research methods (Chapter 3) that the same would hold true under NCEES (which is also based on Danielson's [2007] *Framework*). If it has been found that teachers under NCEES have improved (i.e., moved from "proficient" to "accomplished"), it could be inferred that

student outcomes similarly improved; and if teachers have shown no improvement (or were worse) within NCEES, it could be inferred student outcomes were similarly related. This approach drove investigation into whether teachers in the sampled county have improved across the *Framework's* ratings and dictated statistical tests used within the investigation.

Batten et al. (2012) found evidence from 2010-2011 NCEES data that conflicted with that of Milanowski et al. (2004), Heneman et al. (2006), and Sartain et al. (2011). Batten et al. analyzed data stemming from NCEES which has been the teacher evaluation system in the county under study in Chapter 3 of the primary research. Batten et al. found a weak relationship between EVAAS® outcomes and Standard 4 ratings for teachers, noting that Standard 4 directly relates to teaching practices in the classroom. Unlike Milanowski et al. (2004), Heneman et al., and Sartain et al., Batten et al. found that teachers rated higher on Standard 4 by principals did not necessarily provide higher EVAAS® outcomes. Batten et al. relayed the following in relation to this finding:

> Most notably, Teacher Effect data had a relatively weak relationship with teachers' Standard 4 evaluation ratings. In fact, for each standard error that a teacher's individual effect lay above zero, that teacher's Standard 4 rating would increase by about four hundredths of a point on average. To put this in perspective, for the 2010-2011 school year, a teacher rated "Above Average" according to EVAAS® data had an 8 percent chance of having a higher Standard 4 ratings than a teacher who was comparable in every other way, except for an EVAAS® score of "Not Detectably Different." (p. 5)

Milanowski et al. (2004), Heneman et al. (2006), and Sartain et al. (2011) have

provided evidence for making an inference that when a teacher improves in ratings within a teacher evaluation system, student achievement increases; but the TNDOE (2012) study and Batten et al. (2012) suggested caution in making this inference. This was another example of why researchers have been so interested in validating quality teacher evaluation systems–when teachers were rated higher by their principals, it could be inferred that their students were performing at a higher level. Historically, principal judgments about teacher performance have been deemed a subjective practice (Peterson, 2000); but with this new research and evidence, validated teacher evaluation systems should provide a more objective outcome in the future.

Throughout the 1970s, 1980s, and 1990s, it was difficult to cross tabulate and measure the effects of the performance of specific teachers on student achievement because of a combination of inadequate teacher observation and evaluation rubrics and a lack of student outcomes that could be compared to teacher observation ratings; but the decades following have shown promise (Kane, Taylor, Tyler, &Wooten, 2010). The approach of evaluating teachers based on student outcomes (specifically standardized tests) was evolving in the 1980s and the 1990s but was not widespread; there were also few studies linking student outcomes on standardized tests to the quality of teacher observation and evaluation rubrics (Peterson, 2000; Wise et al., 1984).

The NCLB Act of 2002 provided researchers important teacher and student data stemming from mandated standardized tests in reading and mathematics for Grades 3-8 and once during high school for reading, mathematics, and science (New America, 2013). In present-day literature spanning the 2000s and forward, data stemming from the NCLB Act of 2002 has been used to measure the correlation of teacher performance and their effects on student achievement outcomes (Batten, 2013; Kane et al., 2010; Heneman et

al., 2006; Milanowski et al., 2004). These studies concentrated on measuring correlations from observation and evaluation ratings using quality research-based rubrics (Kane et al., 2010).

By measuring the statistical correlation of student outcomes from mandated NCLB Act standardized tests and teacher performance rating outcomes from observation and evaluation rubrics, researchers could judge whether teacher observation and evaluation rubrics were designed with quality and validity. Alongside the policy impetus to improve teacher evaluation systems since the NCLB Act of 2002 and the RttT initiative of 2009 came an improvement in the quality of standardized tests (Daley & Kim, 2010). The combination of higher quality teacher evaluation systems built using research-based approaches and higher quality student assessments led the way to current approaches in using correlative measures to validate quality teacher evaluation systems. The connectivity of an increase in available data, an improvement in the quality of standardized tests, and an improvement in the ability of evaluation systems to make valid judgments concerning teacher performance positively affected teaching practices and teacher leadership by differentiating among teachers based on the quality of their teaching practices and teacher leadership (Kane et al., 2010).

Kane et al. (2010) made use of standardized testing data from the decade following the NCLB Act implementation in Cincinnati, Ohio; and the researchers compared data from student outcomes on the state standardized tests to teacher ratings from the Cincinnati Teacher Evaluation System (CTES) rubric. The goal was to measure the correlation of teacher ratings from CTES to that of student outcomes on the state standardized tests in order to gauge the validity of CTES in increasing student achievement and to measure the quality of teacher observation and evaluation rubrics in

doing so (Kane et al., 2010).

Kane et al. (2010) was significant because of its large sample size that included 2,071 teacher evaluation records over the course of a decade, with most teachers being observed and evaluated many times using the CTES rubric. Kane et al. (2010) was unique compared to the studies of Milanowski et al. (2004) and Heneman et al. (2006), although it shared some similarities in that Kane et al. (2010) measured the quality of the teacher observation and evaluation rubric by measuring the correlation between teacher ratings and student outcomes on standardized tests.

Kane et al. (2010) differed because the researchers investigated specific elements on the teacher observation and evaluation rubric that affected teaching practices and teacher leadership the most and made the largest measurable impact on student learning. This approach allowed Kane et al. (2010) not only to judge the quality of the teacher observation and evaluation rubric (Danielson's *Framework* used within CTES) but to use the rubric to identify specific teaching practices and leadership that improved student learning.

Kane et al. (2010) estimated the relationship between the CTES rubric ratings of teachers and student achievement and found that the higher a teacher's rating using the CTES rubric, the more student achievement increased–this outcome paralleled findings from Milanowski et al. (2004) and Heneman et al. (2006) and later mirrored in Sartain et al. (2011). Kane et al. (2010) also found that disaggregating for the effects of teachers on student learning resulted in a correlation of 0.105 for math and 0.141 in reading based on CTES rubric teacher ratings and as found in the data table in Appendix S. This outcome was apparent while controlling for student characteristics (column 2), teacher experience (column 3), and school characteristics (column 4) (Kane et al., 2010).

In Kane et al.'s (2010) findings (using Danielson's *Framework*), a teacher ranked in the top TES quartile ("distinguished") produced learning gains (on average) of three percentile points in reading and two percentile points in math for students at the 50th percentile, compared to a teacher ranked in the bottom quartile ("unsatisfactory"). While Kane et al. (2010) admitted this increase may not have seemed large, the increase specifically signified the effect as measured by the CTES rubric for only the chosen standards within Danielson's *Framework*. Kane, Taylor, Tyler, and Wooten (2011) noted the following while writing in *Education Next* in retrospect of their 2010 study:

> This difference might not seem large but, of course, a teacher is just one influence on student achievement scores (and classroom observations are only one way to assess the quality of a teacher's instruction). By way of comparison, we can estimate the total effect a given teacher has on her students' achievement growth; that total effect includes practices measured by the TES process along with everything else a teacher does. The difference between being taught by a top-quartile total-effect teacher would be about seven percentile points in reading and about six points in math. This total-effect measure is one example of the kind of "value-added" approach taken in current policy proposals. (p. 59)

Furthermore, Columns 5 and 6 in Appendix S garnered further support of the correlation between teacher ratings on the CTES rubric and student achievement within Kane et al. (2010), alongside the quality and validity of the system. Columns 5 and 6 were a subset of teachers for which Kane et al. (2010) measured the correlation of CTES ratings and student achievement over time by using CTES ratings for multiple years (2 or 3 years). Kane et al. (2010) observed that "while less stable and noisier given the smaller sample," the outcome in Columns 5 and 6 (for Overall Classroom Practices), "remains a

statistically significant predictor of student achievement growth even within teachers" (p. 22).

This was a significant finding because rather than describing a correlation snapshot (as with teachers measured in Column 4 in Appendix S) of teacher ratings and student growth, the teachers measured in Columns 5 and 6 provided a more long-term description of the correlation between teacher ratings from CTES and student outcomes. Teachers with multiple years of classroom observations and ratings (in Columns 5 and 6) provided evidence that these teachers were producing student learning outcomes based on their teaching practices and teacher leadership measured by the CTES rubric rather than some unique characteristics of the teachers themselves (Kane et al., 2011).

The long-term approach of Kane et al. (2011) gave researchers the opportunity to observe effects of teachers over time within CTES rather than snapshot measurements which traditionally defined teacher evaluation systems. Ultimately, Kane et al.'s (2011) approach has the potential in the future to improve teaching practices and teacher leadership within teacher evaluation systems by providing valid long-term data that would be available to teachers in the form of feedback.

Lastly, what made Kane et al. (2010) unique compared to similar studies was the ability to look at particular components (i.e., "Overall Classroom Practices," "Classroom Environment Relative to Instructional Practices," "Questions & Discussion Approach Relative to Standards & Content Focus") of the observation and evaluation rubric and specifically identify how student outcomes correlated with those teaching practices (refer to Appendix S). This allowed principals the ability to identify more specifically what teachers in their district needed to improve upon in order to increase student achievement (Kane et al., 2010). Teaching practices and teacher leadership were positively affected

because teachers were then able to design lesson planning and pedagogy based on the practices from teachers who were effective (Kane et al., 2010).

Kane et al. (2011) remarked,

> From these data, we can also discern relationships between more specific teaching practices and student outcomes across academic subjects. Among students assigned to different teachers with the same Overall Classroom Practices score, math achievement will grow more for students whose teacher is better than his peers at classroom management (i.e., has a higher score on our Classroom Management vs. Instructional Practices measure). We also find that reading scores increase more among students whose teacher is relatively better than his peers at engaging students in questioning and discussion (i.e., has a high score on Questions/Discussion vs. Standards/Content). (p. 59)

In measuring the quality of teacher observation and evaluation instruments, Kane et al. (2010), Milanowski et al. (2004), Heneman et al. (2006), Daley and Kim (2010), and Sartain et al. (2011) built upon approaches used in the 1970s with Hanushek (1972) and Murnane (1975); in the 1980s with Medley and Coker (1987); and in the 1990s with Sanders and Horn (1998). The current body of literature attested to the transition away from using simplistic checklists within teacher observation and evaluation processes to using researched-based outcomes to build quality teacher observation and evaluation structures and using VAM to assess the validity of teacher observation and evaluation rubrics and processes (Danielson, 2007; Marzano et al., n.d.). Sartain et al. noted the following in their study relating to the validity of Danielson (2007): “If the Framework and the value-added indicators are valid, we expect, for example, that the teachers with the highest classroom observation ratings are the same teachers with the highest-added

indicators" (p. 10).

This phenomenon is indeed what has been uncovered within a growing body of current educational literature. Some correlation coefficients have been found to be categorically "moderate" based on Dancey and Reidy's (2004) classification of correlation levels (0: none; 0.1-0.3: weak; 0.4-0.6: moderate; 0.7-0.9: strong; 1: perfect). Researchers have calculated correlation coefficients using an evaluation system's teacher ratings versus value-added data. Such studies included Milanowski et al. (2004) and Heneman et al. (2006) as aforementioned, but those studies averaged to be at the lower end of the "moderate" classification. Other studies have shown similar outcomes with correlation coefficients categorically at a higher "weak" level or lower end of the "moderate" level. Such studies included Holtzapple (2003), Jacob and Lefgren (2008), Daley and Kim (2010), Headden (2011), and Stronge & Associates (2013).

Stronge & Associates (2013) reported their teacher observation and evaluation rubric has had stronger correlational relationships among its seven standards than Danielson's (2007) *Framework* alongside comparable sample sizes. In surveying various studies, this claim has shown to be plausible historically and in the present day, but the correlations lie mostly in the higher "weak" level and in the lower "moderate" range (on average), albeit higher than Danielson's *Framework*. Stronge & Associates' (2013) correlation measurements among standards are listed in Appendix T for reference. Stronge & Associates (2013) involved 338 teachers in the Commonwealth of Virginia and showed moderate levels of correlation (at statistically significant levels) when comparing teacher ratings from Standards 1-6 to student learning (Standard 7). This finding was important because of the higher levels of correlation with respectable sample sizes.

Stronge & Associates (2013) also provided another set of data indicating the correlation of the six standards among themselves and are also noted in Appendix U. Stronge & Associates' analysis of the correlation among the six process standards within their teacher evaluation system was unique and allowed for an analysis of which standards taken together were associated at higher levels. Standard 1 (Professional Knowledge–strong content and pedagogical knowledge) was closely associated with Standard 3 (Instructional Delivery–differentiation and strategic delivery of content) and Standard 6 (Professionalism–successful networking and collaboration).

The significance of this outcome was two-fold. First, Stronge & Associates (2013) was able to show the overlap among standards, meaning that the "impression of the teacher's competence underlies all of the ratings and accounts for the correlation among these standards." (p. 9). Second, Stronge & Associates' approach was important because of the possible link within teacher evaluation systems that would infer a teacher's knowledge of their content area coupled with their pedagogical knowledge being associated with higher ratings based on how they delivered the material to their students and how they interacted with colleagues. Stronge & Associates' overlap was important for adding validity in that teachers who ranked low or high across standards produced parallel outcomes when calculating their correlations with student achievement.

Both sets of data from Kane et al. (2010) (derived using Danielson's Framework in Cincinnati) and Stronge & Associates (2013) substantiated the validity of their teacher observation and evaluation instruments within teacher evaluation systems by measuring student learning. As with most of the correlational studies carried out, both Kane et al. (2010) and Stronge & Associates identified the advantages of measuring this correlation within teacher evaluation systems, and the advantages included using teacher evaluation

systems to identify high-performing teachers for the sake of reward and replication of their teaching practices and teacher leadership. Also, teacher evaluation systems would be able to provide lower performing teachers with professional development opportunities to improve teaching practices and teacher leadership.

VAM and principal ratings are only two indicators available to judge the effectiveness of teachers and ultimately test the validity of teacher observation and evaluation rubrics and processes within teacher evaluation systems. There have been other indicators that have added to the validity of teacher observation and evaluation, such as student and parent surveys, alongside teacher self-assessment surveys (MET, 2012). These indicators have shown to produce positive effects on teaching practices and teacher leadership through their use in teacher evaluation systems by producing added validity and the ability to differentiate among teachers based on performance (Marzano Research, 2011; MET, 2012).

Using multiple measures to evaluate teachers has ensured quality teacher evaluation systems making them more valid in identifying high-performing teachers for replication and low-performing teachers for professional development or dismissal. Forty-one states have required or have recommended using multiple measures as part of their state teacher evaluation systems (Hull, 2013). Multiple measures also aided in defending higher-stakes teacher evaluation systems where chronically low-performing teachers were at risk of dismissal (MET, 2010).

The MET (2010, 2012, 2013) studies emphasized the importance of using student surveys to add validity to teacher evaluation systems and, in doing so, the predictive power of establishing the top and bottom 25% of teachers based on the quality of teaching practices and teacher leadership increased substantially. The MET (2010, 2012,

2013) studies were unique in that they were longitudinal (spanned 3 years), included a large sample size of 3,000 teachers in spread geographical regions, and used a robust system of validity and reliability measures. The MET (2010, 2012, 2013) studies also used panoramic cameras in teachers' classrooms that allowed a feel for the true learning environment and allowed observers to rate teachers without the limitations of the observer effect or halo effect. While overcoming some barriers to observation and evaluation within the teacher evaluation system, MET's (2010, 2012, 2013) approach added another variable with the teacher constantly sensing the chance of being randomly observed.

One feature that the MET (2010, 2012, 2013) studies specified was the careful consideration to randomization. MET (2010, 2012, 2013) used VAM and student survey data from sections of students the teacher had taught and measured the pairwise correlation of VAM outcomes and student survey data (from those sections) to achievement gains in *different* sections (and the prior year) that the same teacher taught. MET (2010) provided their reasoning:

> For example, even among those with similar prior academic achievement (which the value-added measures control for), one group of students may be unusually well behaved. In that section, a teacher's value-added is likely to be positive, but the students' perceptions of the classroom climate and the score of the video observation may also be unusually high. Therefore, if we were to focus solely on the same section correlations, we run the risk of overstating the predictive power of a given measure. In contrast, the different section or prior year correlations are based on data from distinct groups of students, where the teacher is the common factor. (p. 24)

MET (2010) found that although increasing the number of observations within teacher evaluation systems (as was advised in the previous section of this literature review) increased validity, incorporating student survey data increased validity to a higher degree; and MET (2010) found that the adjusted (for error) correlations between the student survey data versus the math value-added data were 0.43 for a different section of students and 0.34 for the prior year of students (reading was 0.18 for a different section of students and 0.20 for a different year of students). Overall, MET (2010) concluded that the various correlation measurements based on teacher observation ratings, VAM outcomes, and student survey data from different sections or the prior year, would increase the validity of teacher evaluation systems to the point that

> combining observation scores, student feedback, and student achievement gains was better than graduate degrees or years of teaching experience at predicting a teacher's student achievement gains with another group of students on state tests, and combining observation scores, student feedback, and student achievement gains on state tests was also better than graduate degrees or years of teaching experience in identifying teachers whose students performed well on other measures. (p. 2)

In identifying struggling teachers, MET (2012) noted that the ultimate goal was to provide feedback to help teachers improve their practice; and by using multiple measures within teacher evaluation systems, principals could then more uniquely and specifically offer struggling teachers professional development opportunities that would improve teaching practices and teacher leadership. In using multiple, validated judgments of teacher performance within teacher evaluation systems, teachers were ensured that judgments were not made superficially and subjectively as in previous modes of teacher

evaluation systems (MET, 2012). Also, in talking about instruction in postconferences, outcomes of teacher evaluation systems were aligned with the necessary suggestions from principals about what would most improve teaching practices and teacher leadership (MET, 2012).

Using multiple measures with teacher evaluation systems, including data from student surveys, VAM estimates, and principal judgments, has been shown to increase the ability to differentiate among teachers based on the quality of their teaching practices and teacher leadership (MET, 2010). It has been shown that this approach supplied principals with better information in order to guide their efforts in improving teaching practices and teacher leadership. For example, MET (2010) provided graphical details on how multiple measures of teaching practices and teacher leadership has helped differentiate among teachers by using data from student surveys coupled with VAM estimates (see Appendix V).

The MET (2010) graph in Appendix V showed how teacher evaluation systems could use VAM and student survey data in conjunction with observation and evaluation ratings to identify the teachers with quality teaching practices and teacher leadership based on the level of evidence available. By adding VAM data from teacher evaluation systems, the gap between the initial evidence collected by the Tripod student survey widened distinguishing among teachers with a greater degree. This differentiation scheme has led to teacher evaluation systems pinpointing teachers with the most effective teaching practices and teacher leadership and studying their characteristics in the hope of replicating them (MET, 2010).

The culminating MET project has added a considerable amount of evidence of the effects of quality classroom observations used within teacher evaluation systems, and the

study has shown the effects of the quality outcomes on teaching practices and teacher leadership. The MET (2012, 2012, 2013) project boasted of being the largest study to date in validating teacher observation and evaluation instruments (the study included 7,491 videos recorded using video technology from Teachscape and tested five different classroom observation instruments for validity); and the study noted the improvement in validating teacher evaluation systems when including multiple data sources (MET, 2012).

Specifically, MET has shown that when using valid classroom observation rubrics within teacher evaluation systems, effective teachers can be identified, unlike the previous historical period where Weisberg et al. (2009) found that teachers were automatically deemed "satisfactory" by superficial and subjective measurements. The ability to identify effective teachers using validated teacher evaluation systems has resulted in the MET project supplying a video library of teachers who have been identified as effective and using their approach for use in professional development. Struggling teachers could then see effective teachers in action and replicate their approach (MET, 2013).

Another important impact from the MET project is the movement away from judging teachers based on credentials rather than the outcomes of their teaching stemming from validated teacher evaluation systems. Many states have moved away from concluding that teachers are effective because they possess a Master's degree; and because of the degree of validity outcomes teacher evaluation systems have recently received, some districts have structured teacher pay based on teacher evaluation outcomes (Turner, 2010**)**.

The validity of a principal's ability to judge the performance of a teacher within teacher evaluation systems has been improving thanks to a large body of research

providing validity to the process using VAM and student feedback as correlates. The improvement in the validity of teacher evaluation systems has been documented in considering correlations ranging from weakly negative (Peterson [2000] cited Brookover [1940] as reporting a -0.078 correlation between student outcomes and principal ratings); to zero in both the 1950s and 1980s studies from Anderson (1954) and Medley and Coker (1987), respectively; and finally to weak and lower moderately correlated outcomes from more recent studies such as Holtzapple (2003), Milanowski et al. (2004), Heneman et al. (2006), Jacob and Lefgren (2008), Kane et al. (2010), Sartain et al. (2011), Curtis (2011), Headden (2011), and Stronge & Associates (2013).

The improvements in correlation outcomes within teacher evaluation systems has been noteworthy because the correlations validated the measures currently being instituted based on expectations from the RttT initiative of 2009 in states across the country. Historically, if teacher evaluation systems were to improve teaching practices and teacher leadership, it was essential that principals had the capacity to collect data from multiple, correlated sources in order to professionally develop those teachers who needed help. That approach was missing in the past but has been revitalized with work such as the MET project.

Recent studies emphasizing correlation sizes have also been important because if quality teacher evaluation systems were to be used in making high-stakes decisions in teacher promotion to leadership positions, awarding teachers continuing employment, rewarding teachers with higher pay based on performance, or used to terminate employment of teachers, the system should display fidelity (Peterson, 2000). If quality teacher evaluation systems were to be used for low-stakes decisions such as determining professional development assignments for teachers, the results should be trustworthy

(Peterson, 2000). Correlation measurements gave credibility to this approach. Milanowski, Heneman, and Kimball (2011) noted the following:

> We would expect to see correlations between assessment scores and value-added estimates in the .2-.5 range. Correlations of this size are meaningful in terms of the long-run improvement of faculties if assessment scores are used for human capital management decisions such as tenure, compensation, and remediation. Smaller correlations are evidence that either the assessment system is not focusing on the most important drivers of student achievement or the measurement procedures are not reliable or being implemented as intended. Calculating these correlations provides districts with important evidence that can be marshaled to justify the use of teaching assessments for human capital management decisions and to improve the assessment system. (p. 24)

Headden (2011) carried out a study of Washington D.C.'s teacher evaluation system (IMPACT) and revealed the following correlation between D.C. teachers' IMPACT classroom observations and VAM outcomes:

> In a perfect world, a high correlation would be .8 or .9. In fact, it is .34. The finding is perhaps not surprising given that tests measure limited competencies, whereas good schools teach a far broader set of skills. Indeed, noting that high correlations are rare in the social sciences, Thompson calls the figure "moderately strong" and "relatively encouraging." (p. 11)

In an attempt to measure the validity of North Carolina's NCEES and its effects on teaching practices and teacher evaluation, Batten (2013) recently calculated a North Carolina statewide correlation for EVAAS® versus teacher evaluation ratings from principals. Batten's analysis also included correlations for district-level samples. This

finding was important because the study involved a large sample of teachers (N=26,260), and EVAAS® has been the ongoing method of VAM calculations for the sample of teachers and principals chosen in this study within the primary research in Chapter 3. NCEES has relied on the ratings of principals from Danielson's (2007) *Framework* outcomes from EVAAS® to supply data for summative evaluation for teachers in North Carolina.

Batten (2013) found a statewide correlation of 0.23 between a composite measure of principal ratings on Standards 1-5 and Standard 6 EVAAS® outcomes. Batten also found a 0.22 correlation statewide between teacher Standard 4 ratings and EVAAS® outcomes. Standard 4 of NCEES was important for Batten because this standard directly related to teacher pedagogy.

Batten's (2013) outcomes were found to be on the lower end of the 0.17-0.49 range, derived from Batten's review of the literature. District-level correlations had a range of -0.14 to 0.61 due to larger errors for sampling in smaller districts (Batten, 2013).

Lynn et al. (2013) provided a contradicting outcome to that of Batten (2013). While Batten relied on correlation measurements, Lynn et al. relied on regression calculations. Also, while both Batten and Lynn et al. used data that overlapped from the 2011-2012 school year, Lynn et al. used EVAAS® outcomes from 2008-2009 to 2010-2011 school years, while Batten used EVAAS® outcomes solely from the 2011-2012 school year. Regardless of these differences, both studies were contradictory in their conclusions, but it was noteworthy that both studies used different methodologies and differing statistical tests.

While Batten (2013) found a low correlation of EVAAS® outcomes in the

literature compared to other VAM models, Lynn et al. (2013) found a regression trend that was "positive and significant for all standards" (p. 26) related to principal ratings versus Standard 6 EVAAS® outcomes, while using Ordinary Least Squares (OLS) estimates. Lynn et al. used 6-10 years of teacher experience and "developing" categories as comparison groups. The data supplied by Lynn et al. are provided in Appendix W for reference.

The significant relationship between principal ratings and EVAAS® outcomes were noted in the Lynn et al.'s (2013) study, and it was important to note Lynn et al.'s outcome contradicted Batten's (2013) correlational study. However, conclusions from Lynn et al. also contradicted Batten et al.'s (2012) study, which was based on regression using OLS estimates like Lynn et al. This established an apples-to-apples comparison in that both Lynn et al. and Batten et al. were regression measurements using OLS. Batten et al.'s results are noted in Appendices K and L and show lower regression coefficients that were markedly different than the regression coefficients found by Lynn et al. in Appendix W.

While Batten et al. (2012) concluded that Standard 6 EVAAS® outcomes had a "relatively weak relationship" (p. 5) with Standard 4 from NCEES, Lynn et al. (2013) concluded that the relationship between Standard 6 EVAAS® outcomes was "positive and significant for all standards" (p. 26). It was important to note that both Batten et al. (2012) and Lynn et al. used similar sample sizes and data from a similar time range. The Batten et al., Batten (2013), and Lynn et al.'s studies are revisited in Chapter 3 in order to compare these outcomes with the perceptions of teachers and principals about the validity of EVAAS® and teacher ratings using the *Framework*.

Validating principal judgments involving teacher performance has a bright future given the interest and data available, especially in light of the expectations of the RttT initiative in revamping teacher evaluation systems (Marzano et al., n.d.). One common theme validity studies have shared has been the ability to identify which teachers need help, and valid teacher evaluation systems are structured and naturally enabled to provide this ability (Danielson, 2007; Marzano et al., n.d.; Stronge & Associates, 2013).

Historically, it was difficult to identify which teachers increased student achievement, and this subjectivity made it difficult to offer professional development to teachers who needed it the most. With newly revamped teacher evaluation systems, not only should struggling teachers be identified more easily, but because the rubrics used to rate teachers have a valid research base lending credibility, struggling teachers should be able to pinpoint target areas on the rubric for which they scored low and concentrate on improving their performance in those areas (Daley & Kim, 2010; Kane et al., 2010; Sartain et al., 2011).

Reliability within teacher evaluation systems. Validity studies involving teacher evaluation systems have increased with the impetus of the RttT initiative, and the statistical reliability of teacher evaluation systems consistently has been measured alongside validity outcomes (Sartain et al., 2011). Validity has been defined as the ability of an instrument to measure what it is intended to, and reliability has been defined as the ability of an instrument to produce repeatable outcomes within acceptable parameters (Sartain et al., 2011; Stronge & Associates, 2013). Reliability has had an important impact on teacher observation and evaluation systems as it has given credibility to the processes involved which has led to quality outcomes (Little, Goe, & Bell, 2009; MET, 2012). Statistically, “reliability is a necessary precondition of validity” (Cohen et al.,

2007, p. 133).

Historically, the reliability of teacher evaluation systems has been under scrutiny alongside validity, and reliability has been low due to inflation of ratings for various reasons including a lack of training (MET, 2013), a fear of hindering morale (Collins, 2014) and the "observer effect" and "halo effect" (McIntyre, 1980; Medley & Coker, 1987). When judgments among observers have not been reliable, it has led to the denigration of the quality of teacher evaluation systems, and validity then became questionable (Sartain et al., 2011). Also, if VAM outcomes or student survey data have not been reliable, the validity of teacher evaluation systems has become suspect (MET, 2013; Stronge & Associates, 2013). Reliability has helped build the case for validity (Cohen et al., 2007).

Evidence has been mixed in the current literature in terms of the reliability for teacher evaluation systems, but evidence has shown that reliability has increased with newly designed teacher evaluation systems spurred by the RttT initiative (Little et al., 2009). Historically, Soar et al. (1983) relayed the reliability involving the achievement gains of students for particular teachers and revealed that several studies at that time estimated the median reliability of mean student gains of teachers at 0.30. Soar et al. also mentioned that measurement experts advised a reliability of at least 0.90 when making decisions about employment, keeping in mind that the 0.30 value reported was the median value of the dataset available at that time.

Stronge & Associates (2013) included examples of reliability which were "important for establishing an objective and valid evaluation system" (p. 17). Criterion-related reliability referred to the closeness of which a trained rater agrees with an expert rater (such as a trainer); inter-rater reliability referred to the closeness of ratings from two

or more observers who were similarly trained; finally, intra-rater reliability referred to the closeness of ratings from an observer to an external observer.

Like validity requirements, reliability measures should meet accepted statistical levels. Stronge & Associates (2013) identified a reliability level of 0.70 or higher as adequate and anything less than 0.50 inadequate; although as aforementioned, Soar et al. (1983) recommended a reliability measurement of at least a level of 0.90 when making employment decisions.

These recommendations have fallen in line with George and Mallery (2003) where if the reliability using Cronbach's alpha was $\alpha \geq 0.9$, the comparison between raters was excellent for high-stakes decisions; if $0.7 \leq \alpha < 0.9$, the comparison between raters was good for low-stakes decisions; if $0.6 \leq \alpha < 0.7$, the comparison was acceptable; if $0.5 \leq \alpha < 0.6$, the comparison between raters was poor; lastly, if $\alpha < 0.5$, the comparison between raters was unacceptable.

Stronge & Associates (2013) pointed out higher reliability measurements produced by an analysis of their Virginia study alluded to in the previous validity subsection of this chapter. Strong & Associates also outlined the positive effects of their teacher evaluation system on teaching practices and teacher leadership as a result of higher reliability levels.

Not only did an analysis of the validity yield better-than-expected results given comparable studies, the results of inter-rater reliability tests using Stronge & Associates' (2013) rubric produced highly reliable outcomes as referenced in Appendix X. The percentages given indicated the number of raters who agreed with expert raters in Stronge & Associates.

Intra-rater reliability was also found to be higher using Stronge & Associates'

(2013) rubric among principals within 24 schools and among multiple classroom observations (in the Virginia study), yielding an average value of reliability of 0.83 among observers at multiple schools. Stronge & Associates' findings in this area are also included in Appendix Y for reflection.

Because Stronge & Associates' (2013) rubric has shown strong inter-rater and intra-rater reliability, it provided confidence in its validity and allowed for quality feedback for teachers in that observers were confident in what they were judging. This approach gave principals and teachers specific areas provided by evidence stemming from multiple observers, multiple observations, and data provided by VAM that teachers used to improve in their planning and delivery of instruction (Stronge & Associates, 2013).

Sartain et al. (2011) included details about reliability in their study of CPS, which were already alluded to in the validity subsection of this chapter. Specifically aforementioned, Sartain et al. reinforced the validity of Danielson's (2007) *Framework* by providing evidence for a correlative distinction between higher performing teachers who produced high VAM outcomes and lower performing teachers who produced low VAM outcomes.

In terms of testing reliability outcomes, Sartain et al. (2011) included data from trained, external observers who independently observed teachers in their classroom, and the authors calculated the statistical odd ratios that principals would provide reliable observations by comparing their outcomes to the independent observers. This approach coincided with Stronge & Associates' (2013) definition of intra-reliability. The goal was to analyze the reliability of Danielson's (2007) *Framework* in CPS and the effects of reliability on teaching practices and teacher leadership.

Sartain et al. (2011) found that principal ratings (N=4,747) aligned with the external observers (N=4,852) especially at the lower "unsatisfactory" and "basic" ratings, but the results showed discrepancies at the higher "proficient" and "distinguished" levels. Sartain et al.'s data are provided for reference in Appendix Z.

Sartain et al. (2011) found that principals and external observers agreed on the bottom and middle of the scale where the odds ratio was not significantly different, meaning the principals and external observers agreed. However, Sartain et al. found that at the higher levels, especially the "distinguished" level, principals rated teachers with much higher outcomes (6.18 odds ratio of principals versus external raters and statistically significant at the 0.1 level).

When Sartain et al. (2011) controlled for principals' previous classroom observations by matching "distinguished" teachers who had previously been identified as such, "much of the variation between principal and observer ratings disappeared" with an odds ratio of 1.50 that was not statistically significant (p. 16). Sartain et al. also noted that although the principals tended to rate more teachers as "distinguished" than the external observers, VAM outcomes agreed with the principals in that those teachers produced higher student achievement outcomes. By comparing ratings from principals to VAM, Sartain et al. reinforced the use of using multiple measures to ensure validity and reliability.

Wise et al. (1984) provided historical evidence that although struggling teachers search for ways to improve teaching practices and teacher leadership, effective teachers also search for ways to improve teaching practices and teacher leadership using feedback from principal observations and evaluation. Sartain et al. (2011) provided evidence that principals were better at reliably identifying the bottom performers with some

discrepancies in distinguishing among top-tiered performers.

Regardless of the discrepancy of the principals' ability to rate top-performing teachers as reliably, Sartain et al. (2011) indicated that meaningful discussions about teaching practices and teacher leadership still took place in postconferences; and there was an improvement in teaching practices and teacher leadership because of the discussions. High performing teachers alongside average and low-performing teachers received crucial feedback that helped improve teaching practices and teacher leadership across the spectrum of teacher performance, and this outcome was specifically attributed to being a beneficial result of high reliability in observing and evaluating teachers (Sartain et al., 2011).

The principal's ability to make reliable judgments about the performance of teachers in a classroom setting was reinforced with the studies provided by Sartain et al. (2011) and Stronge & Associates (2013). With increased reliability came targeted feedback from principals to teachers about how they could improve specific teaching practices and teacher leadership roles while increasing student learning (Sartain et al., 2011; Stronge & Associates, 2013). Rather than arbitrary and subjective teacher ratings found historically within teacher evaluation systems (Peterson, 2000), the approach of Stronge & Associates and Sartain et al. utilized reliable teacher ratings based on evidence that were shown to increase student achievement and correlate with multiple measures of teacher effectiveness.

An historical shortcoming of teacher evaluation systems was the lack of training raters received producing poor reliability outcomes (Brandt, Mathers, Oliva, Brown-Sims, & Hess, 2007). Brandt et al. (2007) examined a large sample of teacher evaluation systems in midwestern school districts (N=140 school districts) and found that 8% of the

systems involved provided training for raters. This lack of training has caused raters to be inconsistent in how they approached using rubrics and assigning ratings, which resulted in lower reliability measurements (Brandt et al., 2007). Other studies produced similar outcomes (Appeldoorn, 2005; Curtis & Wiener, 2012).

The work of Appeldoorn (2005) and Curtis and Wiener (2012) questioned the ability of a principal within teacher evaluation systems to be able to appropriately judge teacher performance with reliability and professionally develop teachers when principals were not appropriately trained to identify areas in which teachers needed improvement. When teachers were not appropriately advised on areas of improvement because of a lack of reliability in teacher ratings, there were negative effects on teaching practices and teacher leadership (Appeldoorn, 2005; Curtis & Wiener, 2012).

Contrary to this approach, when sufficient, in-depth training is provided to raters within teacher evaluation systems, higher levels of reliability ensue and teachers are more likely to improve teaching practices and teacher leadership (Curtis, 2011; Daley & Kim, 2010; Donaldson & Peske, 2009; Heneman et al., 2006; McCullough, 2011; Piburn & Sawada, 2000). Daley and Kim (2010) cited using multiple observers including the certification and yearly recertification of those observers as the main drivers of high reliability within classroom observations and evaluations among schools using the TAP evaluation system. When there were discrepancies between trainee observers and expert raters within TAP training sessions, trainee observers received extensive training until appropriate levels of intra-rater reliability were reached (Daley & Kim, 2010).

The results of high reliability among multiple observers in positively affecting teaching practices and teacher leadership were noted in reviewing the literature. Headden (2011) indicated that teachers tended to internalize reliable rubrics with which they were

judged, and the teachers consistently concentrated on demonstrating behaviors associated with the rubrics because of the reliable feedback they received. This was a favorable approach with the important caveat that the rubric used to judge the teachers was validated, reliable, and replicable (Headden, 2011).

Furthermore, MET (2013) noted that when school districts gave careful attention to the reliability of teacher evaluation systems, teachers sought to understand the observation and evaluation rubric in order to gauge the effectiveness of their own teaching practices and teacher leadership. This paralleled Headden's (2011) insight.

Jerald (2012) recommended a "calibration" process where teachers were given "opportunities to reach a deeper understanding of the observation instrument so they can begin to calibrate their own vision for effective practice against it" (p. 24). Highly reliable outcomes have peaked the interest of teachers to be more knowledgeable about the observation rubric used to observe and evaluate them, which has led to an improvement in teaching practices and teacher leadership in order to meet the expectations of the rubric (Headden, 2011; Jerald, 2012).

Historically, McGreal (1983) noted that rubrics used to rate teachers within teacher evaluation systems were not up to the task in that they did not measure validated and reliable teacher performance that was associated with student growth. Little et al.'s (2009) review of teacher evaluation systems reinforced McGreal (1983). Little et al. found that the observation and evaluation rubrics historically used to judge the classroom performance of teachers were narrowly and unreliably designed to capture a small subset of teacher behaviors associated with student learning. However, Little et al. noted an improvement in standards-based observation and evaluation of teachers in recently designed teacher evaluation systems to provide reliable outcomes and meaningful

feedback to improve teaching practices and teacher leadership.

Producing a validated, reliable, and replicable rubric to be used within teacher evaluation systems has been the goal within the educational community. As Jerald (2012) indicated, when teacher evaluation systems unreliably rate teachers, the misclassification can lead to inconsistency in accurate feedback given to teachers. This could lead to "significant opportunity costs" (p. 3) in that struggling teachers may not be given accurate feedback to improve, and effective teachers may be given misguided feedback that could cause confusion (Jerald, 2012).

Ho and Kane (2013) questioned the ability of a principal to offer the correct amount and type of meaningful professional development if struggling teachers were rated inconsistently across multiple observers. Jerald (2012) stressed the importance of accurate feedback from principals and showed how feedback from principals within the Classroom Assessment Scoring System (CLASS) boosted student achievement on Virginia's state tests from the 50th to 59th percentile (on a 100-point scale). However, Jerald also contrasted the effects of unreliable feedback from principals to teachers and how such an outcome negatively affected teaching practices and teacher leadership.

The most prevalent negative effect of inaccurate feedback was the misclassification of teachers as aforementioned (Jerald, 2012). Jerald (2012) was concerned about missing the opportunity to improve student achievement when potentially struggling teachers were not identified and offered sound feedback or professional development. The students of struggling teachers ultimately had deficient learning outcomes with lower achievement and growth (Jerald, 2012).

Recommendations from the literature advised a "calibration" process when teacher evaluation systems have provided unreliable teacher ratings in displaying low

inter-rater or intra-rater reliability (Daley & Kim, 2010; Henry, 2011; Joe, Tocci, Holtzman, & Williams, 2013; Little et al., 2009; MET, 2012). Little et al. (2009) described calibration as

> Calibration refers to a periodic assessment of whether raters are continuing to score reliably. Raters trained to use a certain instrument may "drift" from their original training. For example, there is a tendency for raters to score teachers differently at the beginning of the year compared to later in the year–after observing more teachers, they may become more lenient or more stringent in their scoring. This potentially results in teachers receiving different scores for the same performance. Valid evaluation systems will protect against this rater "drift" by establishing rater reliability not just at the beginning of the process but periodically throughout the year and will provide continued training to recalibrate raters to reliable levels. (p. 23)

Teacher ratings from observers who are consistently calibrated led to higher reliability among teacher performance ratings within teacher evaluation systems and provided validity to outcomes (Little et al., 2009). According to Little et al. (2009), teachers have perceived the reliability of outcomes when receiving classroom observations and evaluations from observers, and if those observation ratings fluctuated among observers, the "utility and credibility" (p. 7) of feedback was threatened. Without credibility stemming from reliability, teachers would not be in the position to accept feedback from teacher evaluation systems and teaching practices and teacher leadership would not improve (Headden, 2011; Jacques, 2013).

Teacher evaluation systems yielding low reliability brought about by a lack of calibration hindered the improvement of teaching practices and teacher leadership in the

past (Goldstein & Noguera, 2006; Little et al., 2009). When observers were appropriately calibrated, the outcomes of reliable feedback led to an improvement in teaching practices and teacher leadership; and reliable feedback also allowed more concrete personnel decisions at the district level in terms of the retention and dismissal of teachers (Goldstein & Noguera, 2006).

Henry (2011) explored the advantages of large-scale calibration techniques on a large sample of administrators (N=2,093) using the CLASS observation and evaluation rubric. Henry also analyzed the effects of the CLASS rubric on the reliability of ratings after calibration and the effects of professional development on teaching practices and teacher leadership stemming from CLASS feedback from administrators. This feedback was an impetus for calibration efforts seen as a potential driver of the improvement of teaching practices and teacher leadership within teacher evaluation systems (Henry, 2011).

Henry (2011) was significant because it showed the ability to calibrate a large sample of observers and measure reliability using Cronbach's alpha. Henry found that 71% of raters reached acceptable reliability levels after the first phase of assessment. Henry elaborated that if this level of reliability was replicated in school districts, a "workforce of calibrated raters" (p. 140) would be produced. Most importantly, Henry found that this level of reliability resulted in the ability of administrators to identify teaching practices and teacher leadership that were associated with student learning outcomes. Henry also found that based on the reported reliability levels, administrators would have an increased ability to confidently offer professional development opportunities to improve teaching practices and teacher leadership.

Similar to Henry (2011), Hamre, Pianta, Mashburn, and Downer (2007) relayed

the calibrated efforts of highly trained raters to observe over 4,000 teachers using the CLASS rubric and the effects of those observations and evaluation on teaching practices and teacher leadership. One goal of Hamre et al. was to show the potential of calibrating rating instruments within teacher evaluation systems and the effects on teaching practices and teacher leadership. In Hamre et al.'s study, raters were highly trained in sessions where they were calibrated by observing and rating teachers based on video recordings, and the raters were mandated to pass an intra-reliability test by attaining an 80% match to master codes. From the classroom observation data, Hamre et al. calculated internal consistencies using Cronbach's alpha ranging from 0.77 to 0.89, which were considered acceptable levels by the authors.

Hamre et al. (2007) concluded with the following remark relating the success of CLASS as a reliable source of data within the processes associated with teacher observation and evaluation:

> Using the CLASS observational tool as the metric for the quality of teacher-student interactions, Pianta and colleagues provided teachers with feedback on their interactions using a combination of video and text, with results indicating that ongoing engagement in a cycle of feedback produced significant gains in the quality of teachers' interactions with students for dimensions reflective of each of the three broad domains of interaction tested in this investigation. (p. 19)

Hamre et al. also advised further study using the CLASS system to specifically identify the characteristics of teaching practices and teacher leadership that led to the improvement of teacher performance so that the results could be used to train other teachers using professional development.

In an attempt to increase the reliability of classroom observations, teacher

evaluation systems have been evolving away from using a single observer and evaluator to observe and evaluate teachers, such as a school-based administrator (Ho & Kane, 2013). One example of such an approach is the Peer Assistance and Review (PAR) teacher evaluation system (Goldstein & Noguera, 2006).

PAR has shown to increase the reliability of teacher evaluation outcomes within teacher evaluation systems by allowing consulting, peer teachers to be major stakeholders in the observation and evaluation process (Goldstein, 2004). The increased reliability outcomes from PAR evaluation systems stemmed from using consulting teachers who served between 1 to 4 years (Humphrey, Koppich, Bland, & Bosetti, 2011), were chosen based on their expertise in the classroom (Goldstein & Noguera, 2006), and were released from their classroom duties to perform as observers (Humphrey et al., 2011). These consulting teachers then served full-time in observing and evaluating beginning teachers or veteran teachers referred to PAR by their principal (Humphrey et al., 2011).

In a study by Humphrey et al. (2011), consulting teachers worked one-on-one with participating PAR teachers for an average of 80 hours during a school year in conjunction with an average of 190 pages of detailed paperwork chronicling the performance of the teacher participating in the PAR evaluation system. Humphrey et al. noted the targeted, reliable feedback that consulting teachers were able to provide teachers and the improvement that resulted from using the PAR evaluation system. Consulting teachers and principals could recommend to an overseeing panel that participating PAR teachers exit the PAR program, continue in the PAR program until they reach an expected level of improvement, or be dismissed (Humphrey et al., 2011).

Relative to the approach used in the PAR system, Ho and Kane (2013) provided evidence that increasing the number of classroom observers within teacher evaluation

systems increased the reliability of teacher ratings markedly. Ho and Kane showed the improvement in reliability by using two raters to observe and rate two lessons ("1 → 2 Raters–2 lessons") for the same teacher rather than using one rater to observe and rate one lesson ("Reliability of 1 Rating by 1 Observer"), one rater to observe two lessons ("1 → 2 Lessons–1 rater"), or two observers rating a lesson for the same teacher ("1 → 2 Raters–1 lesson"). Ho and Kane's data are in Appendix A1 for reference.

Ho and Kane (2013) was significant because it showed that teacher evaluation systems using multiple samples of observations and ratings increased reliability more when adding an additional observer ("1 → 2 Raters–1 lesson") rather than increasing the number of observations and ratings from a single observer ("1 → 2 Lessons–1 rater"). Specifically, Ho and Kane noted, "The gain in reliability from adding another set of eyeballs is more than twice as large as that of adding another observation from the same observer" (p. 22). This extra "set of eyeballs" (Ho & Kane, 2013, p. 22) has been intrinsic to the approach in the PAR system in that consulting teachers provided added reliability, and the extra reliability provided detailed evidence and targeted feedback to beginning and struggling veteran teachers participating in the PAR system (Goldstein & Noguera, 2006; Humphrey et al., 2011).

MET (2012) found similar results as Ho and Kane (2013) in adding more observers in order to increase reliability. MET (2012) found a reliability of about 0.65 across four of five of the rubrics tested using four observers. For further reference, the results are shown in Appendix B1 as supplied by MET (2012).

MET (2012) specifically found that there was an increase in the ability of observers to capture accurate judgments of teaching practices and teacher leadership when multiple observers were utilized and statistical noisiness was reduced. This enabled

teachers to receive targeted feedback on ways to improve teaching practices and teacher leadership (MET, 2012).

In summary, evidence supporting the validity of teacher evaluation systems has been historically problematic because of the lack of multiple measures of teacher effectiveness available to researchers. Historically, studies that were carried out showed little connection between the performance teachers and student learning (Coker et al., 1980; Soar et al., 1983).

As a result, teacher evaluation systems were rarely validated, and their outcomes were considered perfunctory (Weisberg et al., 2009). The effects of teacher evaluation systems on improving teaching practices and teacher leadership were historically limited, mostly as a result of the inability to appropriately identify and distinguish among teachers based on their performance (Danielson & McGreal, 2000; Peterson 2000).

Recently revamped teacher evaluation systems have been formed around performance standards that have been shown to correlate with student learning. Rubrics from Danielson (2007), Stronge & Associates (2013), and Marzano et al. (n.d.) have been validated using statistical methods. Historically, teacher evaluation systems relied on checklists that may have been agreed upon by state and local officials but were rarely validated based on student outcomes. The newly validated teacher evaluation systems have created a valuable distribution of teacher performance within a spectrum of performance standards rarely available to researchers and policymakers in the past (Batten et al., 2012; Curtis, 2011; Daley & Kim, 2010).

The distribution of teachers has given principals the ability to identify teachers at the low and high ends of teacher performance with more fidelity than ever considering the history of teacher evaluation systems described by Coker et al. (1980) and Soar et al.

(1983). This has been the key ingredient to reforming teacher evaluation systems– identifying low and middle performers and offering them professional development opportunities to improve teaching practices and teacher leadership while identifying high performers for the sake of learning from them and replicating their teaching practices and teacher leadership.

Recently revamped teacher evaluation systems employing multiple measures of teacher effectiveness have enabled correlation studies uncovering the statistical relationship between subjective principal judgments and objective student learning stemming from indicators such as VAM. Danielson's (2007) *Framework* took on special interest in this review of the literature because it has been the instrument used within NCEES since its implementation. Studies from Milanowski et al. (2004), Heneman et al. (2006), Kane et al. (2010) and Sartain et al. (2011) showed a weak to moderate level of correlation between that of Danielson's *Framework* and student outcomes using VAM as a measure.

The second part of this subsection concentrated on reliability of teacher evaluation systems in order to uncover its effects on teaching practices and teacher leadership. After distributing teachers based on performance and validating those outcomes based on multiple measures, the literature concentrated on measuring the degree of precision (i.e., reliability) the validated outcomes provided.

The reliability of teacher evaluation systems to offer credible feedback has been under scrutiny historically. Soar et al. (1983) estimated the median reliability of mean student gains of teachers across several studies at 0.30, which was a weak reliability outcome.

There has been evidence provided by Sartain et al. (2011), Stronge & Associates

(2013), and MET (2010) that the reliability of newly revamped teacher evaluation systems has improved. The improvement in the reliability of teacher evaluation systems has aided in spurring conversations among teachers and principals about the quality of teaching practices and teacher leadership (MET, 2010; Sartain et al., 2011). The more reliable the feedback has been from principals, the more credibility the feedback carried; and teachers improved their teaching practices and teacher leadership proportionally (Sartain et al., 2011; Stronge & Associates, 2013).

Furthermore, multiple authors showed the importance of training raters in order to increase reliability outcomes from teacher evaluation systems which in turn provided more exact details for which teachers could use to improve teaching practices and teacher leadership (Goldstein & Noguera, 2006; Henry, 2011; Ho & Kane, 2013; MET, 2010). Lastly, research has shown that multiple observations were important for increasing the reliability of teacher evaluation systems, but using different observers increased reliability to a greater extent (MET, 2012).

To summarize, teachers in pilot data revealed little about their perceptions of the validity of NCEES or other teacher evaluation systems they had experienced in the past. However, many teachers expressed a level of agreement in the outcomes of their past experiences with being evaluated, and some provided vague accounts of how principals helped them improve teaching practices and teacher leadership. Specificity was missing in their accounts. It was inferred that teachers generally thought past evaluations they received were valid, although they offered few specifics as to how their perceived validity affected their teaching practices and teacher leadership.

One reason why few details from teachers may have been offered in this arena was that the focus groups to collect pilot data were carried out in the summer of 2012.

NCEES had just rolled out, and many teachers expressed difficulty in logging on to the online portion of NCEES. Other than math teachers in focus groups, teachers had little input about their perceptions of EVAAS®, but teachers also had not experienced NCEES fully operational. It was not clear whether teachers were provided EVAAS® outcomes or how many received EVAAS® for the courses they taught. This made it difficult to validate the observation and evaluation process involved in NCEES with EVAAS® outcomes.

The reliability of feedback was a concern among several teachers in separate focus groups. Two respondents in different focus groups expressed similar accounts that the outcomes of their evaluations were directly related to whichever administrator was carrying out the classroom observations and final evaluations. Both of these teachers and teachers as a whole agreed that the feedback they received was helpful, even if the quality varied by administrator. No details about how or why the feedback was helpful was offered. Responses in this regard seemed superficially generic.

The second category, "Quality of Classroom Observations and Teacher Evaluation," was inferred initially from pilot data supplied by administrative interviews and teacher focus groups. The category was derived from coding and analyzing the pilot data using a grounded theory approach as outlined by Charmaz (2006). Further investigation ensued spanning literature from multiple decades, and those data combined with data from the grounded theory analysis were used to construct and refine the categories depicted in Figure 2.

This methodological approach resulted in the secondary research methods found within this literature review, and the data reported in this secondary research were used to

drive the primary research methods found in Chapter 3 of this project. Specifically, given the theme "Quality of Classroom Observations and Teacher Evaluation," the data collected from continued administrative interviews and a teacher survey (during the primary research methods in Chapter 3) were compared against data from this secondary research in the Literature Review in an effort to verify parallel conclusions.

The Effects of Using VAM

Validity of VAM. A reoccurring theme throughout the literature has been the pursuit of using teacher evaluation systems to differentiate among teachers based on their performance (Goldhaber & Hansen, 2010; Sanders & Rivers, 1996; Weisberg et al., 2009). The goal in using teacher evaluation systems to differentiate between teachers has been to enable top performers to be rewarded and their practices be replicated, to allow for specific professional development to be offered to low and middle performers, and to enable the dismissal of consistent low performers (Weisberg et al., 2009).

Historically, this differentiation among teachers based on performance has not generally taken place–the majority of teachers have been found to be good or great, with very few teachers found to be anything less (Weisberg et al., 2009, p. 10). The undifferentiated results from teacher evaluation systems have brought their validity and reliability into question and have been perceived as hindering the improvement of teaching practices and teacher leadership (Peterson, 2000).

As was observed in the preceding section entitled *The Effects of Teacher Evaluation Systems–Quality of Classroom Observations and Teacher Evaluation*, teacher ratings within teacher evaluation systems have historically been shown to be inflated (Peterson, 2000; Soar et al., 1983). Because of the inherent issues of the "observer effect" or "halo effect" (McIntyre, 1980; Medley & Coker, 1987) within rating schemes of

teacher evaluation systems, the fear of hindering teacher morale (Peterson, 2000), or the fear of receiving union grievances (NCTQ, 2011b), teacher ratings tend to be inflated towards higher outcomes (Weisberg et al., 2009).

In order to improve upon the subjective outcomes (Peterson, 2000) of teacher evaluation systems and largely due to the RttT initiative, a more quantitative and objective approach known as VAM has been instituted (Goe, 2008). As Goe (2008) asserted, "observer bias can be minimized but not eliminated; with value-added models, there is no observer–only scores" (p. 5).

The scores used within VAM alluded to by Goe (2008) have originated largely from state-mandated standardized tests and have been used by teacher evaluation systems using VAM to estimate the contributions of teachers to the learning growth of their students (Sanders & Rivers, 1996). Results from the ACT standardized test have also been used to drive school-level accountability using VAM (Sanders & Horn, 1998; State Collaborative on Reforming Education [SCORE], Taking Note, 2012). VAM has used various standardized test results by scaling scores, but it has been recommended that the test have "a strong relationship to the curriculum" (Sanders & Rivers, 1996, p. 2).

VAM has been used within teacher evaluation systems to estimate the effectiveness of teachers based on the value the teacher has added to the academic growth of a student between two or more points in time stemming from scores on standardized tests (Bianchi, 2003; Hershberg, 2005; Kupermintz, 2003). VAM has calculated a student's predicted score based on the student's performance on past tests, and VAM then estimated value-added outcomes for a teacher's effect by comparing a student's recently observed score to the student's predicted score (SAS, 2007). If a student's observed score was higher than the predicted score, the student displayed positive academic growth, and

the teacher was considered to have positive effect as measured within a teacher evaluation system (SAS, 2007). If a student's observed score was lower than the predicted score, the student displayed negative academic growth and the teacher was considered to have a negative effect as measured within a teacher evaluation system (SAS, 2007).

Some VAM models used within teacher evaluation systems control for student background variables and project predicted scores using minimal previous test scores (McCaffrey et al., 2004), while other VAM models used within teacher evaluation systems do not directly control for student background variables as covariates and use at least 3 years' worth of test scores to serve as intrinsic controls for students (Sanders, Wright, Rivers, & Leandro, 2009). By using 3 years' worth of test data, students serve as their own control because their SES status and other characteristics are intrinsically built into their historic test scores (Sanders et al., 2009).

Using VAM estimates of teacher effectiveness within teacher evaluation systems has transformed teaching practices and teacher leadership in that teachers concentrate on showing academic growth among their students rather than simply concentrating on moving students to the proficient level (Cody et al., 2010; Wei, Hembry, Murphy, & McBride, 2012). Under the NCLB Act of 2002, schools and teachers were held accountable for moving students to proficient levels on state-mandated, standardized tests; but data stemming from the NCLB Act of 2002 was rarely incorporated within teacher evaluation systems (Hershberg, 2005, Wei et al., 2012). Teachers potentially concentrated on improving the achievement of students who could move to proficient levels, perhaps to the detriment of the lowest achieving students who were perceived has being too low and too difficult to move to proficient levels (Hershberg, 2005; Pearson

Learning, 2004). Those students who were in special education programs or considered at-risk were potentially ignored within this scheme (Hershberg, 2005; Pearson Learning, 2004).

The implementation of VAM within teacher evaluation systems has been meant to "level the playing field" (McCaffrey, 2013, p. 2) because teachers would be held accountable to show growth across a spectrum of diverse learners with potentially different capabilities and characteristics. Rather than concentrating on the demand to make students proficient, teacher evaluation systems using VAM have recently concentrated on students' academic growth regardless of the student's current status in regards to proficiency (Hershberg, 2005). Even students who have been far from being proficient could still show growth, and newly revamped teacher evaluation systems using VAM have attempted to fairly take this into account (Hershberg, 2005).

In terms of the validity of VAM used within teacher evaluation systems and how validity issues tie into improving teaching practices and teacher leadership, VAM has calculated estimates for the effectiveness of teachers with the ability to control for a myriad of external variables that have affected student learning (Goe, 2008). Teachers have been able to observe an estimate of their ability to facilitate student achievement due to the implementation of VAM within teacher evaluation systems (Sanders & Rivers, 1996). Rather than exclusively relying on subjective feedback from principals, teacher evaluation systems have been coupled with VAM to give teachers the ability to measure their effectiveness based on a more objective measure (Goe, 2008). Validity studies involving VAM have helped give teacher evaluation systems credibility in judging teaching practices and teacher leadership (Goe, 2008).

The validity of teacher evaluation systems incorporating VAM has hinged on

their ability to disentangle the many variables that affect student learning (Darling-Hammond, Amrein-Beardsley et al., 2012). Goldhaber (2002) estimated 60% of those variables involved factors outside the school (individual and family background characteristics) versus approximately 9% of the variables involving differences in teacher effectiveness. Teacher evaluation systems have been revamped using VAM to produce an estimate of teacher effectiveness based on the growth of their students with some flexibility for error; and to measure the validity of that aim, researchers have disaggregated VAM outcomes against variables that are known to correlate with student learning (Darling-Hammond, Amrein-Beardsley et al., 2012; Wei et al., 2012).

Teacher evaluation systems incorporating VAM have accounted for variables such as the socioeconomic status (SES) of students and various peer characteristics (ethnicity, gender, etc.) at the district and state levels (Sanders & Rivers, 1996). In order for VAM to have met statistical validity demands and to have been used fairly within teacher evaluation systems, it should not have significantly correlated with students' SES, which has been an important indicator of poverty (Darling-Hammond, Cook et al., 2012).

Since the Coleman Report of the 1960s, there has been debate among researchers and policymakers as to what extent teaching practices and teacher leadership have affected student outcomes and to what extent teachers can overcome extraneous variables outside the environment of the school (Hershberg, 2005). When used within teacher evaluation systems, VAM has shown the ability to successfully identify effective teachers amid the myriad of variables that affect student learning without showing a significant correlation with students' SES (Sanders & Rivers, 1996). Sanders and Rivers (1996) indicated early in the history of VAM that a goal was to incorporate VAM within teacher evaluation systems to uncover teaching practices and teacher leadership of effective

teachers and find out how these effective teachers operate.

This approach has placed newly revamped teacher evaluation systems in a new scientific status in that evaluation systems are not just rating teacher performance. With the infusion of VAM and linking teacher ratings to teacher VAM outcomes, teacher evaluation systems are front and center in determining what teaching practices and teacher leadership are causing the greatest growth in student achievement.

Teachers have also been positively affected by teacher evaluation systems using VAM because it has given teachers the ability to measure the quality of their teaching practices and teacher leadership on low achieving versus high achieving levels of students and modify their approaches if necessary (Sanders & Rivers, 1996). If teachers have experienced high VAM outcomes with high achieving students but their low achieving students failed to show growth, teachers have had the ability to reflect on what may work better to reach these students and modify their approach in the future (Sanders & Rivers, 1996). Teacher evaluation systems in conjunction with VAM have given teachers this opportunity.

The underlying theme has been to use VAM to supply teacher evaluation systems with an objective form of data that differentiates among teachers based on performance, which can be used to professionally develop struggling teachers, dismiss chronically low-performing teachers, and reward high-performing teachers (Goe, 2008; Gordon et al., 2006). There also has been a theme in the literature of using VAM related outcomes within teacher evaluation systems to prod teachers to improve using the best possible teaching practices and teacher leadership while showing academic growth among students (Goe, 2008). If teachers do not show expected growth within the framework of VAM, they could run the risk of being found ineffective in conjunction with other

measures within teacher evaluation systems (Glazerman, Loeb, Goldhaber, Staiger, Raudenbush, & Whitehurst, 2010).

Glazerman et al. (2010) reinforced the infrequent number of teachers historically dismissed because of ineffective teacher evaluation systems that lacked the ability to differentiate among teachers based on their performance. The general theme has been that using VAM in a valid and reliable form interjects an objective judgment of teacher performance into teacher evaluation systems; and on that basis, teacher evaluation systems could influence teachers to either improve their practices and leadership or face receiving low VAM growth from their students (Glazerman et al., 2010).

Because valid and reliable VAM outcomes within teacher evaluation systems have estimated the effectiveness of teachers compared to either teachers in their same school, district, or state (McCaffrey, 2013), VAM outcomes have been designed to drive improvement in teaching practices and teacher leadership by inspiring teachers to show large learning gains among their students (Goe, 2008). The goal for teachers has been to strive to fall within Sanders and Rivers' (1996) top quintile of teacher effectiveness, which has led to financial awards and other merit-based rewards in some cases (Amrein-Beardsley & Collins, 2012). Bock, Wolfe, and Fisher (1996) reinforced this idea before widespread implementation of VAM when they remarked,

> To illustrate how the estimated teacher gains might be used, suppose one wished to identify with 95 percent confidence those teachers that are below the 20-th percentile or above the 80-th percentile of the teacher gains distribution. If the gains measure has stability coefficient of 0.80, the lower and upper cut points would have to be set at the 8^{th} and 92-nd percentiles, respectively, in order to insure 90 percent confidence that a teacher was correctly classified in the lower

> and upper 20 percent of the teacher population. In other words, the system would single out eight percent of the teachers as meritorious and eight percent of teachers as problematic with respect to their 3-year average gains. (p. 50)

Sanders and Horn (1998) also reinforced that teachers would seek to display as much academic growth from their students as possible under TVAAS® (the form of VAM used in Tennessee):

> TVAAS reports, issued annually, include information on student gains for each subject and grade for the three most recent years as well as the three-year average gains. The cumulative average gain is the primary indicator by which success is measured. (p. 250)

The competitive nature of VAM within teacher evaluation systems has allowed teachers to observe where they are ranked in performance measures (Goe, 2008) among their peers in contributing to the learning of their students (Cody et al., 2010). Glazerman et al. (2010) recognized the ability of VAM to measure the performance of teachers within this competitive context:

> Teachers and their mentors and principals stand to gain vast new insight if they could see the teachers' performance placed in context of other teachers with students just like their own, drawn from a much larger population than a single school. This is the promise of value-added analysis. (p. 4)

Using VAM within teacher evaluation systems has shown parallels to that of a business model where managers seek to measure the output of workers quantitatively in order to differentiate among employees, all in the hope to drive improvement (Eichenwald, 2012). The Microsoft Corporation used such an approach for almost a decade beginning in the early 2000s, and the approach became known as "stack ranking"

(Eichenwald, 2012). In stack ranking, there were a set number of employees who would be evaluated as low, middle, and top performers. The hope was that by incorporating such an approach as stack ranking in the evaluation systems of Microsoft, employees would naturally seek to improve performance in order to avoid consequences and/or receive recognition and rewards for high performance (Eichenwald, 2012). Teacher evaluation systems that employ VAM have offered such an approach in the education sector where teachers compete with each other in an attempt to improve teaching practices and teacher leadership.

Sanders and Horn (1998) elaborated upon the competitive nature of employing VAM to differentiate among low- and high-performing teachers within teacher evaluation systems. According to Wright et al. (1997), the ultimate goal would be to seek out the characteristics of high-performing teachers and infuse their teaching practices and teacher leadership into elements of teacher evaluation systems:

> It is recognized here, however, that there were no direct, systematic observations of the quality of teaching and learning at the classroom level in this study. Thus, identifying teachers that clearly get results over time, and comparing them to teachers over time who do not, seems a logical, worthwhile next step in addressing the issues raised here and in further developing general lines of inquiry about the important relationship between teacher effectiveness and teacher evaluation. If characteristics of teaching and learning environments that differentiate teachers who are demonstrably effective (as opposed to ineffective) in different contexts over time can be documented, subsequent teacher evaluation systems might be developed to accommodate these characteristics. (p. 66)

While teacher evaluation systems incorporating VAM have been instituted at the

teacher level, school-wide VAM outcomes have also estimated the amount of student achievement growth that has occurred among schools and school districts (Preston, 2013; SCORE, 2012). VAM outcomes at the school level have been termed "school-wide growth," and in some states school-wide growth has been counted for a certain percentage of a teacher's evaluation within teacher evaluation systems (NCDPI, 2013; SCORE, 2012).

Teacher evaluation systems coupled the competitive nature of VAM at the teacher level and the collaborative approach at the school level by incentivizing teamwork to increase school-wide growth (NCDPI, 2013; TNDOE, 2012). In Tennessee, school-wide growth has counted for 15% of a teacher's personal evaluation within their teacher evaluation system, while 35% of growth has been based on student growth on the TCAP assessment for individual teachers (TNDOE, 2012). In this capacity, teacher evaluation systems have the ability to institute competitive practices and the ability to differentiate among schools and school districts, based on the quality of teaching practices and teacher leadership within them on an aggregate scale.

This competitive drive within teacher evaluation systems has yielded positive results in Tennessee (TNDOE, 2012). TNDOE collected and analyzed substantial quantitative data from state-mandated tests and robust qualitative data from teachers and administrators and published the results in a report already partially analyzed in the preceding subsection of this literature review, *First to the Top–Teacher Evaluation in Tennessee: A Report on Year 1 Implementation* (TNDOE, 2012).

Within the TNDOE (2012) study, an unprecedented improvement in student proficiency and growth rates was noted, and the TNDOE (2012) study credited the new teacher evaluation system as a significant reason for the improvement based on data

received from teachers and administrator surveys. Student test scores in Tennessee "improved, in aggregate, at a faster rate than any previously measured year" including "55,000 more students at or above grade level in math than in 2010" and "38,000 more students are at or above grade level in science" (TNDOE, 2012, p. 3).

The TNDOE (2012) study noted that "teacher observation results from year one are encouraging and demonstrate more meaningful differentiation than ever before" (p. 4). The "differentiation" (TNDOE, 2012, p. 4) within the newly revamped teacher evaluation system stemmed from TVAAS® outcomes that were spread across quintiles of performance in a more equal proportion than the qualitative judgments from principals as observed in the table in Appendix J.

While this research may have brought principal ratings from teacher evaluation systems into question as analyzed in the previous subsection of this chapter, TVAAS® has shown to successfully differentiate among teachers based on student outcomes in considering the quality of their teaching practices and teacher leadership (TNDOE, 2012). According to state education officials in Tennessee, principals in Tennessee have given assistance and feedback to struggling teachers who have transformed teaching practices and teacher leadership leading to their successful gains because of the new teacher evaluation system and its incorporation of TVAAS® data (TNDOE, 2012).

Similar to the focus on using principal ratings from teacher observation rubrics to differentiate teachers based on their performance (Curtis, 2011; Daley & Kim, 2010; Sartain et al., 2011), VAM has had this focus in common within teacher evaluation systems (Daley & Kim, 2010). For instance, similar to the TNDOE (2012) study, Daley and Kim (2010) compiled the aggregate VAM outcomes distribution for states using

EVAAS® within the TAP teacher evaluation system; and the results are included in Appendix C1 for review.

Daley and Kim's (2010) analysis and graph were significant because within the TAP teacher evaluation system it involved the distribution of a large sample of teachers (N=7,500), reflected the level of instruction given to a large sample of students (N=85,000 students), and supported the differentiation of teachers across a wide spectrum of performance. Daley and Kim's rating outcomes were unlike historic teacher observation and evaluation outcomes such as in Weisberg et al. (2009) where the majority of teachers were skewed towards higher ratings. Also, because EVAAS® has been the VAM framework used in NCEES at the time of this study, Daley and Kim carried significant inferences. Does the county in which the primary research for this study share a similar distribution of EVAAS® scores? Such a question is posed in Chapter 3 of this study in order to investigate the use of EVAAS® in NCEES in differentiating among teachers based on performance.

While the distinct differentiation among teacher performance (Daley & Kim, 2010; TNDOE, 2012) has been an important issue to the validity of VAM and its effects on teaching practices and teacher leadership, the largest looming issue in the literature has been the ability of VAM to disentangle student-learning outcomes from the variables that affect student learning. If VAM outcomes happen to be correlated with students' SES status, the validity of VAM and the teacher evaluations that incorporate it could lose the credibility of leveling the field for teachers. The ability of VAM to provide teachers with their effects on student learning and positively affect teaching practices and teacher leadership could also be questioned as a result. In the future, VAM could run the risk of

incurring the same hesitancy as using proficiency levels to judge the performance of teachers and their teaching practices and teacher leadership–it would be viewed as a lens to judge the quality of students being taught rather than judging the effectiveness of a teacher. The possible correlation of VAM with student poverty measures could affect the ability of teacher evaluation systems to operate and provide teacher ratings with credibility.

Sanders and Rivers (1996) showed the effects of teaching practices and leadership from effective teachers on large samples of students using TVAAS® data in the 1990s and measured the ability of TVAAS® to overcome students' SES. Sanders and Rivers was significant because it displayed the potential of using VAM as a part of a teacher evaluation system and the ability to differentiate among teachers by estimating their influence on their students using growth outcomes from Tennessee's TCAP achievement test. By dividing teachers based on performance quintiles and tracing the progression of students assigned to these teachers, Sanders and Rivers determined that higher performing teachers produced mean learning gains of at least 50 percentile points when students were assigned to those higher performing teachers for 3 years in a row, rather than being assigned to the lowest performing teachers for 3 years in a row. Moreover, Sanders and Rivers showed that the achievement levels or ethnicity of students did not affect VAM outcomes, and that VAM was a valid judgment of teacher effectiveness as a result.

Sanders and Rivers (1996) noted the advantages school administrators would gain in collecting TVAAS® data within teacher evaluation systems and how teaching practices and teacher leadership alongside student outcomes could be improved within teacher

evaluation systems. Such teacher evaluation systems would give administrators the capacity to use TVAAS® (or EVAAS® in North Carolina) data to institute professional development opportunities targeting struggling teachers to improve effectiveness or choose a teacher sequence that would avoid students receiving lower performing teachers for 3 years in a row (Sanders & Rivers, 1996; SAS, 2007).

A number of studies alongside Sanders and Rivers (1996) have tested the validity of VAM outcomes by examining growth patterns compared to variables that have been shown to closely correlate with student learning, of which the SES of students has been the most noteworthy (Chetty et al., 2011; Kane, McCaffrey, Miller, & Staiger, 2013; Kane & Staiger, 2008). Proficiency levels of state-mandated tests along with other standardized tests have shown to closely correlate with students' SES status; and Hershberg (2005) argued that teachers who serve advanced students had an advantage when proficiency levels were used to judge the quality of teaching practices and teacher leadership. Teachers who taught advanced and gifted students have been much more likely to reflect higher proficiency rankings among their students but not necessarily because of teacher performance as much as the economic level of the parents of the students they teach (Hershberg, 2005).

Chetty et al. (2011) showed VAM as a viable data source in teacher evaluation systems while being uncorrelated with variables affecting student learning. The authors used financial data from family tax records of students representing a significant sample size (n=2.5 million) to show that VAM was uncorrelated with students' SES (Chetty et al., 2011). Top-performing VAM teachers produced learning outcomes for students who led to being more likely to attend college and higher-ranked colleges, to earn $250,000 more lifetime earnings on average, to live in higher SES neighborhoods, and to save more

for retirement (Chetty et al., 2011). Perhaps the most important finding in this study was that VAM outcomes showed the advantage of high-performing VAM teachers without being significantly correlated to students' SES status (Chetty et al., 2011). Chetty et al. added to the validity of VAM as a result of the uncorrelated outcomes between VAM estimates of teacher effectiveness and parent characteristics and showed the potential worthiness of identifying high-performing VAM teachers within teacher evaluation systems.

Chetty et al. (2011) also tested the potential validity of VAM within teacher evaluation systems and the effects on teaching practices and teacher leadership by measuring the change in VAM outcomes for exiting teachers with high VAM outcomes versus low VAM outcomes. It was found that when teachers with high VAM outcomes exited a school, VAM measures decreased accordingly; and when teachers with low VAM outcomes exited a school, VAM measures increased accordingly (Chetty et al., 2011). This result built validity for using VAM with its ability to measure teacher effectiveness within teacher evaluation systems based on the quality of teaching practices and teacher leadership (Chetty et al., 2011).

Other noteworthy studies have supported the ability of teacher evaluation systems using VAM to disentangle variables that affect student learning, while yielding valid and reliable estimations of teacher effectiveness (Bock et al., 1996; Ceperley & Reel, 1997; Glazerman et al., 2010; McCaffrey, 2013; Rivken, Hanushek, & Kain, 2005). An important theme from VAM-related studies was to ensure the validity of VAM while incorporating VAM data within teacher evaluation systems as a tool to hold school systems, schools, and teachers accountable for their performance (Bock et al., 1996).

Ceperley and Reel (1997) outlined possible penalties and/or rewards for schools

and school systems that did not show adequate student growth using VAM outcomes, such as being placed on a probationary status or receiving recognition. Glazerman et al. (2010) dealt with the possible use of teacher evaluation systems incorporating VAM and other teacher effectiveness measures to determine which teachers to lay off in a time of budget cuts and which teachers to dismiss based on inadequate performance. Furthermore, Goldhaber and Hansen (2010) showed that although VAM outcomes were modestly correlated from year to year (inter-temporal stability), VAM outcomes were parallel to data measurement used for high-stakes personnel decisions (i.e., dismissals, layoffs, promotions, bonus pay) in employment sectors outside education and could be used to distinguish which beginning teachers should receive tenure.

A study in Ohio recently analyzed teacher evaluation systems in Ohio and wrought questions about a correlation between VAM and students' SES. This study was significant because it contradicted data supplied by the SAS Institute (O'Donnell, 2013). Although the SAS Institute found little to no correlation between EVAAS® outcomes and students' SES, O'Donnell questioned this conclusion. O'Donnell found that while analyzing data from 1,720 schools in Ohio, of 1,035 teachers in wealthier schools, 34% of teachers were ranked at the top with high growth; while of the 2,411 teachers in poorer schools, 9% were ranked at the top with high growth (see Appendix D1). O'Donnell analyzed EVAAS® outcomes during the 2011-2012 school year and provided the following insights:

- Value-added scores were two and a half times higher on average for districts where the median family income is above $35,000 than for districts with income below that amount.
- For low-poverty school districts, two-thirds had positive value-added scores–

scores indicating students made more than a year's worth of progress.

- For high-poverty school districts, two-thirds had negative value-added scores–scores indicating that students made less than a year's progress.
- Almost 40% of low-poverty schools scored "above" the state's value-added target, compared with 20% of high-poverty schools.
- At the same time, 25% of high-poverty schools scored "below" state value-added targets while low-poverty schools were half as likely to score "below."
- For both districts and schools, high poverty was defined as those with 75% or more students qualifying for free or reduced-price lunches. Low-poverty schools had 25% or fewer of its students qualifying.
- Students in high-poverty schools are more likely to have teachers rated "least effective"–the lowest state rating–than "most effective"–the highest of five ratings. The three ratings in the middle are treated by the state as essentially average performance.
- In the highest-poverty districts, with 90% or more of students qualifying for free or reduced-price lunches, students were twice as likely to have a least effective teacher as a most effective one.
- But the lowest-poverty schools–with less than 10% qualifying for free or reduced-price lunch–had seven times as many most effective teachers as least effective ones.

O'Donnell (2013) asserted that while VAM outcomes were associated with the level of poverty for schools, VAM outcomes "do not rise with income or fall with poverty rates as much as it does with other academic measures . . . but . . . students–and consequently teachers–in richer schools still score better on value-added than those in

poorer schools" (para. 14). The "other academic measures" related to the affluence of families and how SES correlated with proficiency levels of students, of which O'Donnell claimed had "followed that pattern for years" (para. 11).

Along these same lines, Xu (2000) carried out a study that provided evidence that TVAAS® outcomes depended upon the FRL rates of their students. Although Xu was designed to measure the effects of per-pupil-expenditure on student achievement (or proficiency) and student growth (using TVAAS®), Xu uncovered significant findings relating to VAM and family SES for students.

Xu's (2000) findings were significant for several reasons. First, Xu used a robust sample size of 138 school systems in Tennessee. Second, Xu ran Pearson correlation tests uncovering TVAAS® strength of the association between teacher TVAAS® classifications and FRL rates. Third, Xu used regression calculations to predict the level of effects for various independent variables (including students' SES) as predictors for TVAAS® outcomes.

Xu (2000) found significant correlations between the SES status of students and TVAAS® outcomes. Xu also found negative trends in the SES status of students and TVAAS® outcomes using regression calculations. Xu (2000) concluded with the following:

> Value-added scores tend to reduce the initial difference in student achievement levels. In Sanders' words, value-added scores would level the playing field for educators (Sanders, 1993, 1994). Sanders believes that the value-added score blocks the effects of extraneous factors that appear to affect student achievement by using the student as his own control. He even denies the effect of student

> ability on student value-added scores. He thinks his value-added model can statistically filter out external influences so that "you are left with influences that teachers have control of" . . . In this sense, value-added scores may mitigate the influence of those external factors, but they cannot eliminate the influence of those external factors completely. The data in this study shows that student family socioeconomic status has a strong influence not only on student achievement levels but also on gain scores, especially mathematics gain scores. (p. 91)

The importance of the validity of VAM within teacher evaluation systems has rested on the principle of fairness in using VAM outcomes to rate teacher performance. The fear has resided in the idea that teachers would be held accountable for variables out of their control, which could hurt stakeholder buy-in from administrators and teachers (Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2011; Peterson, 2000). It has been shown in the past that if teacher evaluation systems have not been credible, it has had a negative effect on teaching practices and teacher leadership because teachers did not understand the process under which they worked or thought it impossible to meet expectations of unknown standards (Peterson, 2000).

The Wake County Public School System (WCPSS, 2009) Evaluation and Research Department in North Carolina carried out a relevant study to determine the level at which EVAAS® outcomes correlated with student poverty rates at the school level. WCPSS (2009) was interested in whether high-poverty schools and teachers in Wake County were being penalized for teaching high-poverty students. WCPSS calculated their own growth index ("effectiveness index") using a regression formula similarly used within EVAAS® but that controlled for student characteristics including

free and reduced lunch (FRL) status (EVAAS® does not directly contain controls for FRL status).

WCPSS (2009) then compared their growth index regression to that of EVAAS® using correlation measurements and then measured the correlation for both the FRL status of students in schools within the WCPSS growth index regression and the FRL status of students in schools within EVAAS® regression. WCPSS (2009) compiled data related to this topic located in Appendix E1.

WCPSS's (2009) analysis revealed that EVAAS® outcomes for schools correlated significantly with the FRL rates of those schools, meaning high-poverty schools were more likely to receive low growth ratings. Using the WCPSS index regression calculation, WCPSS found no significant correlation with FRL rates. To summarize, WCPSS concluded with the following statements:

> Correlation between the percentage of FRL students in the school and WCPSS Effectiveness Indices were all non-significant, indicating no evidence of a relationship between effectiveness and school poverty, which is not unexpected since school poverty is already factored into those effectiveness calculations. EVAAS® scores, however, were in many cases significantly negatively correlated with school poverty, implying schools with higher percentages of FRL students received lower EVAAS® scores than schools with smaller percentages of FRL students. (p. 4)

Furthermore, WCPSS (2009) showed that for school classification disagreements, Standard 6 EVAAS® ratings tended to classify schools at extreme ends of the distribution ("Low," N=31) ("High," N= 91) otherwise found to be average-based on the WCPSS

regression measurements. This conclusion was drawn from the data also presented in Appendix F1.

More importantly, where school-level Standard 6 EVAAS® classifications deviated, WCPSS (2009) found that a majority of schools (78%) with less than 30% FRL benefited (22% schools with more than 30% FRL benefited); whereas using the WCPSS regression measurements, 68% of schools with more than 30% FRL benefited (32% of schools with less than 30% FRL benefited). WCPSS supplied these data as shown in Appendix G1.

The outcomes from Xu (2000), O'Donnell (2013), and WCPSS (2009) have the potential of negatively impacting teaching practices and teacher leadership in schools where improvement has been needed most (Baker et al., 2010). WCPSS was especially significant because the county under study in Chapter 3 is in the same state and has used Standard 6 EVAAS® classifications for both schools and teachers.

If schools are found to be at a disadvantage in measuring school-level growth, this outcome could thwart school improvement efforts when schools or teachers experience a sense of "false positives" and "false negatives" based on the schools where they teach (Schochet & Chiang, 2010; WCPSS, 2009). WCPSS (2009) specifically alluded to these effects on teaching practices and teacher leadership at the school level:

> Depending on the intended purpose of the measurement–to motivate overall school improvement, to prioritize school improvement goals, etc.–the choice of benchmark may be important, as the cost of potential "false positives" or "false negatives" may depend upon the data user's perspective. (p. 8)

Goldhaber and Theobald (2013) compared VAM formulas used within teacher evaluation systems across the U.S. shown in Appendix H1, of which some included

covariates to control for poverty level of students and others that did not. The results showed that advantaged students showed more growth using the various models, some at significant levels as shown in Appendix I1. The point of Goldhaber and Theobald was to test the correlation of poverty levels for students versus varying forms of VAM formulas used within teacher evaluation systems.

Goldhaber and Theobald (2013) noted that some school districts might have chosen VAM formulas that correlate with student poverty levels because officials have not wanted to send the message they expect any less growth from disadvantaged students than they would want from advantaged students. The concern has been that by choosing to control for poverty, this approach within teacher evaluation systems incorporating VAM would hinder the quality of teaching practices and teacher leadership in that less advantaged students would not be expected to show the same level of growth. Goldhaber and Theobald called it a "judgment call" in determining which VAM formula best suits a school district for use in teacher evaluation systems.

The important conclusion from Goldhaber and Theobald's (2013) analysis was that VAM has shown to correlate with student poverty, at least to some extent. However, as Glazerman et al. (2010) pointed out, it has been impossible for any teacher evaluation system to be perfect. Historically, the NCLB Act of 2002 determined accountability measurements using proficiency levels of students on assessments, which was highly correlated with the poverty levels of students (Hershberg, 2005). Glazerman et al. argued that although VAM outcomes are imperfect, VAM has been the best measurement technique available in teacher evaluation systems and ultimately will help improve teaching practices and teacher leadership.

Another noteworthy outcome of Goldhaber and Theobald's (2013) analysis was

the smaller correlation outcomes for within-school VAM measurements compared to student poverty levels. Goldhaber and Theobald showed when VAM involved teachers from the same school as the baseline for growth calculations rather than district or state data, the poverty levels of students were not correlated with VAM outcomes.

This could be an important development going forward for the fairness of the process of using VAM within teacher evaluation systems to inform teachers of the quality of their teaching practices and teacher leadership. By using within-school VAM rankings, VAM could generate a more valid technique by ensuring teachers they would be compared to teachers in similar circumstances, specifically teaching similar students and operating in similar working conditions. This approach could also afford teachers the ability to improve teaching practices and teacher leadership by zeroing in on processes within their school that would improve student growth.

Whether VAM outcomes correlate with poverty or whether teachers in high-poverty schools have been less effective has been an important question that has arisen, and O'Donnell (2013) mentioned this point of contention has been up for debate. While Sanders and Rivers (1996) and Chetty et al. (2011) were robust studies using large sample sizes and showing VAM to be not statistically correlated with extraneous variables that affect student learning, there have been studies that contradicted these findings to some extent, namely O'Donnell (2013), WCPSS (2009), and Goldhaber and Theobald (2013).

Other studies showing limited to no correlation between the poverty status of students and teacher evaluation systems using VAM were Ballou, Sanders, and Wright (2004), Ballou (2005) and Sass, Hannaway, Xu, Figlio, and Feng (2010). These studies were similar in methodology and involved respectable sample sizes. Ballou et al. (2004)

and Ballou (2005) were relevant because North Carolina uses the same VAM structure within the teacher evaluation system for the county involved in this study.

Sass et al. (2010) specifically found that teachers in high-poverty schools were "generally less effective than teachers in lower-poverty schools" among teacher evaluation systems (p. 22). However, Sass et al. found the differences to be small and inconsistent across the dataset from North Carolina and Florida. Sass et al. was significant because it used large sample sizes in North Carolina (N=9,212) and Florida (N=14,052); and it showed that the correlation between VAM outcomes and the poverty level of students' parents was minimal.

The exception was in high-poverty schools where the least effective teachers skewed the results (Sass et al., 2010). Sass et al. (2010) found that the most effective teachers in high-poverty schools produced VAM outcomes on par with the most effective teachers in low-poverty schools, but the least effective teachers in high-poverty schools skewed the average VAM outcomes of high-poverty schools towards being ineffective.

Both Ballou et al. (2004) and Ballou (2005) showed that the VAM model in Tennessee (TVAAS®) was not affected significantly by student poverty measures when adjusting the TVAAS® formula to account for students' SES status. Even at the extreme ends of the teacher performance where teachers received the highest and lowest TVAAS® outcomes, these teachers were not significantly displaced from their rankings within teacher evaluation systems when considering their outcomes before and after controlling for poverty (Ballou, 2005; Ballou et al., 2004). In Ballou et al. (2004), correlation measurements between the adjusted and unadjusted TVAAS® model exceeded 0.90; while in Ballou (2005), correlation measurements exceeded 0.84. Both studies ultimately

concluded that based on the high degree of correlation between the adjusted and unadjusted models, teacher evaluation systems could successfully employ VAM without directly controlling for student SES as it was inherently controlled for by using 3 years' worth of historical test data. As Ballou (2005) noted, the hope for the future is that by employing VAM within teacher evaluation systems, teachers would improve teaching practices and teacher leadership using data stemming from VAM related outcomes.

One important study offering significant insight was a recent, robust study from Lauen, Henry, Rose, and Kozlowski (2013) which sampled a range of N=363,824 observations of student test scores in mathematics to N=421,498 for reading and included 6,200 fifth-grade teachers; 2,100 sixth-grade teachers; and 1,950 seventh-grade teachers. This study was significant because of its large sample sizes, its findings were recent, and because the study calculated the connection between EVAAS® scores and student and school characteristics that could have been used within NCEES (Lauen et al. data were from the 2009-2010 school year before EVAAS® was incorporated into NCEES). The student characteristics included controlling for minority status, poverty status, and prior achievement; while the school characteristics included controlling for the class average prior achievement and school percentage of poor students. The outcome of Lauen et al.'s study is supplied in Appendix J1 for analysis.

In all four models, Lauen et al. (2013) found a negative association in mathematics for minority and poor students with the regression outcomes significant at a minimum of $p<0.5$ level for minority students and all four models showing the regression outcomes for poor students significant at the $p<0.001$ level. The findings in reading were similar for minority and poor students (Lauen et al., 2013).

Although Lauen et al. (2013) framed their study based on the access that North

Carolina students have had to effective teachers, the study provided evidence that EVAAS® outcomes were associated with student characteristics; in turn, meaning NCEES had the capacity of using this data within the system for Standard 6. Whether the teachers have been ineffective, or that the poverty status of students was a main driver, has been an ongoing debate as was alluded to within the aforementioned O'Donnell (2013) analysis. Lauen et al. offered this insight in assessing the quality of teachers compared to student characteristics (such as poverty status):

> Minority, poor, and low-achieving students typically had lower value-added teachers than did non-minority, non-poor, high-achieving students. However, once classroom and school-level variables are added to the model, results suggest that individual students' characteristics matter less than a classroom's average level of prior achievement. Our findings suggest that schools tend to group students of similar achievement level together and then assign the highest value-added teachers to the classes of students with the highest levels of prior achievement. (p. 3)

Along the same line of research methodology, Newton, Darling-Hammond, Haertel, and Thomas (2010) examined data from 250 teachers and roughly 3,500 students in the San Francisco Bay area in order to measure the correlation of VAM outcomes within teacher evaluation systems to students' SES. This study was significant because the authors devised five models (M1-M5) each controlling for different variables including student characteristics such as a student's SES, and the authors tested the correlation of those models against each other and against student characteristics. The approach paralleled that of Lauen et al. (2013), although Newton et al. looked at correlation outcomes versus Lauen et al.'s regression analysis. The results of Newton et

al. (2010) are placed in Appendix K1 for analysis.

Specifically, Newton et al. (2010) found that teachers who taught a large proportion of ELL students, students eligible for free or reduced lunch, or Hispanic students may have had been disadvantaged in their VAM ranking compared to teachers who did not. Alternatively, teachers who taught a large proportion of Asian students or students with parents who were highly educated may have had an advantage in their VAM ranking (Newton et al., 2010). These results were standard regardless of the VAM model used (Newton et al., 2010)

Newton et al. (2010) also carried out a novel design to test the effects of students being placed on an "upper-track" or "lower-track" on VAM outcomes within teacher evaluation systems. Within Newton et al.'s dataset, the researchers were able to collect and control for teachers who taught both "upper-track" and "lower-track" students within the same time period the data was collected–upper-track students were "more advantaged students" (p. 18) who were placed in a class before their formal allotted time.

However, Newton et al. (2010) relied on a small subsample of three teachers to address this question with specificity, but the results did provide insight and the invitation for further research in this area in the future. Newton et al. displayed data specifically for three teachers (T1, T2, and T3) all of whom taught upper-track and untracked or lower-track classes and compared both tracked classes to VAM outcomes in deciles (see Appendix L1). Untracked classes in Newton et al. were lower-achieving students.

Newton et al. (2010) observed that for upper-tracked students, teachers ranked between the seventh and ninth decile; whereas for untracked students, teachers ranked between the first and third decile.

Newton et al. (2010) concluded with the following remark:

> These findings suggest that teachers who were teaching greater proportions of more advantaged students may have been advantaged in their effectiveness rankings, or that more effective teachers were generally teaching more advantaged students. Our data set allowed us to test this possibility with a set of teachers who taught both upper-track and lower-track courses in the same year. In those analyses we found that students' residualized achievement scores were, in most analyses, more strongly predicted by the students' prior achievement and the course they were in than by the teacher him or herself. Each teacher appeared to be significantly more effective when teaching upper-track courses than the same teacher appeared when teaching lower-track courses. (p. 18)

Two important caveats exist in Newton et al.'s (2010) design and subsequent analysis of data in terms of analyzing the results for teachers who taught upper-tracked and lower-tracked students. First, the sample chosen for the specific investigation was extraordinarily small, and there was no evidence indicating the generalizability of the results for these three teachers to the complete sample of teachers who taught both upper-tracked and lower-tracked students. Second, Newton et al. revealed no methodology in figuring to what extent upper-tracked students were advantaged; rather, it seemed as though it was assumed.

The importance of correlation measurements among VAM outcomes, how they will relate to the validity of teacher evaluation systems, and how these measurements will affect teaching practices and teacher leadership has been the impetus for ongoing research. There has been widespread attention in the literature to the correlation between student proficiency levels and student family poverty levels as measured by SES, usually in terms of FRL levels (Berliner, 2009; Ravitch, 2010; Rothstein, 2004).

Teacher evaluation systems in the past generally did not hold teachers responsible for proficiency levels on state-mandated tests within their personal evaluations (Hershberg, 2005). The NCLB Act of 2002 concentrated more on the qualifications of teachers and holding schools responsible for producing 100% proficiency levels among students by 2014 (Hershberg, 2005).

The approach of holding teachers accountable for the growth of their students framed by VAM has been a turning point in the history and design of teacher evaluation systems, and the validity of newly revamped teacher evaluation systems hinge primarily on whether or not VAM outcomes correlate with students' SES (Johnson, Lipscomb, & Gill, 2013). If it is shown in the future that VAM outcomes correlate with students' SES to some extent and that extent is shown to philosophically or legally skew VAM outcomes out of the control of teachers, this could raise similar concerns as holding schools and teachers accountable for producing high proficiency levels among students (Newton et al., 2010). This topic has been an essential component of the validity of VAM–does VAM measure student growth absent of external influence to an acceptable level, with the understanding that no measure will be perfect (Johnson et al., 2013; Glazerman et al., 2010).

The concerns over correlation and regression measurements involve perceived detrimental effects on improving teaching practices and teacher leadership, especially in light of the high-stake uses of VAM within recently revamped teacher evaluation systems. If correlation and regression measurements involving teacher evaluation systems that employ VAM and students' SES are shown to be consequential in the future, this could draw teachers with quality teaching practices and teacher leadership away from high-poverty urban and rural schools in favor of teaching in more affluent communities

(Newton et al., 2010).

Alternatively, if teacher evaluation systems incorporating VAM were shown not to be correlated with students' SES or at least to a meaningful extent, VAM could provide a crucial reform approach in turning around schools with high SES levels. Turning around the low achievement of high SES schools has been a critical component of the NCLB Act of 2002 and of the RttT initiative of 2009.

After reviewing the literature pertaining to teacher evaluation systems that incorporate VAM, Haertel (2013) summed up the relationship between students' SES and VAM outcomes. After analyzing data involving social stratification, peer effects, random versus nonrandom assignment of students, and summer learning loss, Haertel relayed, "VAM scores *do* predict important student learning outcomes, but my reading of the evidence strongly suggests that these scores nonetheless measure not only *how well* teachers teach, but also *whom* and *where* they teach" (p. 17).

Another concern related to teacher evaluation systems incorporating VAM has been the random versus nonrandom assignments of students to teachers alongside the effect randomization has on teaching practices and teacher leadership (Baker et al., 2010; MET, 2010; Silva Mangiante, 2011). A large concern has been that teachers receive assigned students based on their teaching abilities rather than in a random way; in other words, principals and school counselors may assign effective teachers more difficult students because of their perceived effectiveness (Kupermintz, 2003).

This has the potential of introducing confounding variables and noisy statistical outcomes within VAM structures (Silva Mangiante, 2011) and may also hinder teaching practices and teacher leadership in that teachers may request students who they believe will show the most growth (Ladd & Walsh, 2002). These issues have the potential of

negatively affecting teacher evaluation systems when VAM is employed to measure the effects of students assigned to teachers in a nonrandom distribution (Baker et al., 2010; Briggs & Domingue, 2011; Rothstein, 2008).

Rothstein (2008) highlighted the issues surrounding nonrandom assignment of students to teachers in a data set from fifth graders in North Carolina. Rothstein (2008) used a falsification test and found significant bias in VAM outcomes that would be included in teacher evaluation systems. Koedel and Betts (2009) and Briggs and Dominque (2011) replicated Rothstein's (2008) falsification test and confirmed that using multiple years of data would decrease bias if VAM were used within teacher evaluation systems.

Rothstein (2008), Koedel and Betts (2009), and Briggs and Dominque (2011) all concentrated on testing VAM outcomes by predicting the VAM for future teachers of students while those students were in previous grades. This was a falsification test for validity because, for example, it would be impossible for a fifth-grade teacher to affect VAM outcomes for students in previous grades; yet Rothstein (2008), Koedel and Betts (2009), and Briggs and Dominque all found evidence of this phenomenon.

The outcomes of Rothstein (2008), Koedel and Betts (2009), and Briggs and Dominque (2011) manifested the negative consequences on teaching practices and teacher leadership when teacher evaluation systems employ invalid measures–especially if VAM outcomes used in teacher evaluation systems were publicized (Briggs & Dominque, 2011) as they have been in the past in the Los Angeles Times newspaper. In Rothstein's (2008) view, it was unfair to judge teachers on one specific indicator and over the course of 1 year, especially if that indicator was biased.

Kane and Staiger (2008) tested the validity of VAM within the context of random

versus nonrandom student assignment in the Los Angeles Unified School District and showed that assigning students to teachers randomly versus nonrandomly did not statistically affect VAM outcomes to a significant extent. One limitation of Kane and Staiger was its limited sample size (N=78 pairs of teachers); however, MET (2013) examined the effects of randomization of student assignment on VAM outcomes in a similar methodology to Kane and Staiger and concluded similarly using a larger sample size (n=1,181 teachers) that was spread geographically.

Both Kane and Staiger (2008) and MET (2013) were significant in providing evidence that VAM has estimated teacher effects with valid measures, regardless of whether students are assigned to teachers at random or not. Kane and Staiger and MET (2013) also showed that teacher performance can be differentiated using VAM within teacher evaluation systems regardless of random versus nonrandom student-to-teacher assignment. With the differentiation of teachers producing high versus low VAM outcomes and the confidence that they entail valid measures, teacher evaluation systems have provided development strategies for the improvement of low-performing teachers by providing principals with specified data for professional development planning (Kane & Staiger, 2008; MET, 2013).

While much of the literature has concentrated on the effects of teacher evaluation systems on teaching practices and teacher leadership at the teacher level, the possible correlation of VAM outcomes with extraneous variables that affect student learning and the validity questions that ensue has important implications for studying the effects of teaching practices and teacher leadership at the school level. As Sass et al. (2010) and Goldhaber and Theobald (2013) both detailed, if teachers in more affluent schools are more likely to receive high VAM outcomes within teacher evaluation systems, it is

possible that teachers displaying the most effective teaching practices and teacher leadership could be prone to seeking out and working in such schools. This would have a negative effect on students more macroscopically in high-poverty schools who are naturally in need of receiving effective teaching practices and experiencing positive teacher leadership. Sass et al. advised for policymakers to work to retain teachers producing high VAM outcomes in high-poverty schools while actively recruiting teachers producing high VAM outcomes in low-poverty schools to make the move to high-poverty schools.

The Talent Transfer Initiative (TTI) was designed using federal grant funds (Title II Part A) with the goal of using teacher evaluation systems to retain teachers with high VAM outcomes in high-poverty schools and to recruit teachers with high VAM outcomes in low-poverty schools to transfer to high-poverty schools (Glazerman, Protik, Teh, Bruch, & Max, 2013). The ultimate goal was to use teacher evaluation systems to improve teaching practices and teacher leadership in high-poverty schools using VAM outcomes (Glazerman et al., 2013).

The details of the TTI program included using teacher evaluation systems to offer $20,000 to teachers in low-poverty schools who were ranked in the top quintile of VAM outcomes and who agreed to be compared over the course of 2 experimental years to a control group in high-poverty schools (Glazerman et al., 2013). Teachers ranked by their teacher evaluation system to be in the top quintile of VAM outcomes and already working in high-poverty schools were offered $10,000 for their retention during the experimental treatment years (2009-2011) (Glazerman et al., 2013). The study included a large sample of schools in 10 different school districts with geographical spread.

Glazerman et al. (2013) analyzed data comparing TTI teachers to a control group

and found that the transferring teachers had increased VAM outcomes within teacher evaluation systems that translated into an increase up to 10 percentiles points at the elementary level, which was considered significant. However, there was no impact at the middle-school level (Glazerman et al., 2013). Another upside was that 60% of the transferring teachers found to be highly effective within teacher evaluation systems stayed on after the recruitment payments ended (Glazerman et al., 2013). Teacher evaluation systems successfully aided in the retention of effective teachers in high-poverty schools (Glazerman et al., 2013).

Glazerman et al. (2013) found two relevant conclusions important in building validity for teacher evaluation systems that incorporate VAM and how it affects teaching practices and teacher leadership, especially at the school level. First, teachers who produced high VAM outcomes in low-poverty schools and transferred to high-poverty schools produced equal or better VAM outcomes compared to their counterparts in the control group (Glazerman et al., 2013). This increased the validity of teacher evaluation systems incorporating VAM and their ability to improve teaching practices and teacher leadership in high-poverty schools. Second, Glazerman et al. (2013) showed that it has been possible to retain high-performing teachers in high-poverty schools due in part to a level field of accountability for teachers knowing their effectiveness would be judged more fairly within teacher evaluation systems using VAM rather than using proficiency outcomes.

Using VAM-related outcomes within teacher evaluation systems has allowed researchers and policymakers to investigate two important factors, whether at the classroom level for individual teachers or more macroscopically at the school level. First, it has allowed teacher evaluation systems to make a quantitative differentiation in teacher

performance to isolate successful teaching practices and teacher leadership. Second, in the future, VAM should allow future exploration of the specific teaching practices and teacher leadership roles of teachers producing high VAM estimates within teacher evaluation systems in order to replicate those practices and roles among teachers (MET, 2012).

Reliability of VAM. While validity studies involving teacher evaluation systems using VAM will continue into the future, the reliability of teacher evaluation systems using VAM has also received attention, alongside the effects of reliability on teaching practices and teacher leadership. The validity of VAM within teacher evaluation systems has dealt with whether VAM has measured what it is defined to estimate–teacher effectiveness–not variables associated with student characteristics that affect student learning (Braun, 2005).

The reliability of VAM within teacher evaluation systems has rested in part with statistical measurements made within a school year in order to differentiate among teachers based on performance (McCaffrey, Sass, Lockwood, & Mihaly, 2009). Ballou (2005) carried out a study in which he found that TVAAS® within teacher evaluation systems in Tennessee was able to distinguish 58% of teachers as "significantly different from average at the 10% level" (p. 18) within 1 school year's worth of teacher performance, which he called an "impressive degree of discriminatory power" (p. 18). However, to distinguish 58% of teachers as high or low performers required 2 years of VAM data as supplied by teacher evaluation systems–1 year of data was able to distinguish 30% of teachers because of the margin of statistical error. Bock et al. (1996) and McCaffrey et al. (2009) measured the quality of teaching practices and teacher leadership while both authors reported reliable differences in teacher effectiveness

compared to the mean.

Bock et al. (1996), Ballou (2005), and McCaffrey et al. (2009) had important implications because if teacher evaluation systems were *not* able to use VAM to distinguish enough teachers as performing different from the mean within 1 school year's worth of data, there would be two major repercussions: (1) teacher evaluation systems would not be able to confidently rank-order teachers based on performance; and (2) teacher evaluation systems would not be able to identify bottom and top performers in order to improve the quality of teaching practices and teacher leadership or replicate them in the case of top performers (Ballou, 2005; McCaffrey et al., 2009).

Another important aspect of reliability has been measuring the movement of teachers within year-to-year rankings in quartiles or quintiles–statistically known as intertemporal variation, stability, precision, or generalization (Ballou, 2005; Goldhaber & Hansen, 2010; Haertel, 2013; McCaffrey et al., 2009; Sass, 2008). In this case, authors analyzed movement of teachers among quartiles or quintiles over the course of 2 or more years to determine if movement between divisions was significant or random due to the "noise" of sampling error (Haertel, 2013). In doing so, these authors also revealed details of the effects of the reliability of teacher evaluation systems using VAM on teaching practices and teacher leadership which are interspersed throughout this section.

An important theme in studying the reliability of VAM has been measuring the statistical reliability of VAM in lending support for its validity within teacher evaluation systems and studying the effects of reliability on teaching practices and teacher leadership. This approach was parallel to measuring the reliability of teacher ratings from principal judgments in conjunction with building validity for the observation and evaluation processes within teacher evaluation systems and analyzing the effects on

teaching practices and teacher leadership. As stated previously in that section and equally applicable here, reliability helps build the case for validity (Cohen et al., 2007).

The reliability of teacher evaluation systems incorporating VAM has been important for the same reason the reliability of teacher ratings garnered by administrators using rubrics was important–reliability outcomes have built confidence that teacher evaluations systems have been producing valid results to improve teaching practices and teacher leadership. This approach has built credibility of teacher evaluation systems among administrators and teachers giving administrators confidence in comparing outcomes and giving teachers confidence that their teaching practices and teacher leadership produce measurable outcomes (McCaffrey et al., 2009). In relating validity with a measure of "bias" and reliability with a measure of "imprecision," Ballou (2005) pointed out,

> Although statisticians make much of the distinction between bias and imprecision, it may mean little to the teachers being assessed. Both bias and imprecision imply errors in the estimate of a teacher's effectiveness. In each case, teachers are being held responsible for factors beyond their control. Given the brevity of many teaching careers and the short time horizon for personnel decisions, teachers facing sanctions under a high-stakes value-added assessment system are unlikely to take much consolation from the fact that the system is merely noisy, not biased (if it is even the latter). Teachers will want value-added assessments that are unbiased and precise. Administrators who want to avoid mistaken personnel decisions will want the same. (p. 5)

Some authors have suggested that the year-to-year stability ratings stemming from VAM have fallen within useable parameters, and teacher evaluation systems have

confidently provided data that could be used among multiple measures to make high-stakes or low-stakes decisions as a result (Glazerman et al., 2010; Goldhaber & Hansen, 2010; MET, 2010). In building credibility among stakeholders, teacher evaluation systems have provided reliable VAM outcomes that aided in differentiating among teachers and provided professional development opportunities for low performers (Glazerman et al., 2010; Goldhaber & Hansen, 2010; MET, 2010). In high-stakes use of teacher evaluation systems incorporating VAM, teachers with perpetually low VAM growth could be dismissed, merit pay plans could be implemented, and tenure could be granted based at least in part on VAM outcomes (Glazerman et al., 2010, Goldhaber & Hansen, 2010).

Other authors have argued the year-to-year stability ratings have been questionable and using teacher evaluation systems incorporating VAM to make decisions regarding high-stakes personnel decisions has been detrimental to improving teaching practices and teacher leadership (Amrein-Beardsley & Collins, 2012; Collins, 2014; Darling-Hammond, Amrein-Beardsley et al., 2012; Haertel, 2013). Research institutions such as the American Statistical Association (ASA) and the National Academies Press (NAP) have also presented concerns when using VAM within teacher evaluation systems, especially to make high-stakes personnel decisions for teachers (ASA, 2014; NAP, 2009).

ASA (2014) has been the largest "organization in the United States representing statisticians and related professionals," (p. 2) and ASA expressed caution in using teacher evaluation systems that incorporate VAM for high-stakes purposes. One reason ASA expressed caution was the unstable reliability outcomes of teacher evaluation systems incorporating VAM and the negative impact on teaching practices and teacher leadership. ASA made the following statement:

> The VAM scores themselves have large standard errors, even when calculated using several years of data. These large standard errors make rankings unstable, even under the best scenarios for modeling. Combining VAMs across multiple years decreases the standard error of VAM scores. Multiple years of data, however, do not help problems caused when a model systematically undervalues teachers who work in specific contexts or with specific types of students, since that systematic undervaluation would be present in every year of data. (p. 7)

The previous subsection catalogued the ongoing investigation involving the validity of VAM within teacher evaluation and its ability to disentangle external variables and how teaching practices and teacher leadership have been affected. For example, as presented in the previous subsection, the correlation outcomes comparing VAM outcomes and students' SES were conflicting in the literature. Some studies in the previous subsection showed little correlation while others showed significant correlation of VAM with students' SES. Specifically, the research has centered on the degree of correlation between VAM and students' SES, or the association of teacher VAM outcomes and students' SES using regression, with some studies showing higher levels of correlation and higher regression coefficients than others.

However, researchers in the literature have mostly agreed on the statistical parameters for the reliability of teacher evaluation systems that have used VAM to calculate teacher effects and measure the quality of teaching practices and teacher leadership (Haertel, 2013). It has become a philosophical question as to whether teacher evaluation systems should employ VAM on the basis of its reliability measurements and to what extent outcomes should be used in an attempt to improve teaching practices and teacher leadership (high-stakes versus low-stakes; Goldhaber & Hansen, 2010;

Glazerman et al., 2010; McCaffrey et al., 2009).

For example, two researchers reviewed the literature and found a similar year-to-year reliability range for evaluation systems employing VAM. Goldhaber and Hansen (2010) found a reliability range of 0.20 and 0.60 with most studies producing measures of 0.30 and 0.40 reliability measures. McCaffrey et al. (2009) surveyed the literature and analyzed his own data from Florida, and found a reliability range of 0.2 and 0.40 with a few lower and higher (some middle school outcomes reached the 0.70 level).

Glazerman et al. (2010) cited two meta-analyses justifying the use of VAM within teacher evaluation systems to make high-stakes decisions on the basis of reliability measures and in conjunction with other teacher effectiveness measures. Glazerman et al. (2010) cited Sturman, Cheramie, and Cashen (2005) to show that correlations of worker evaluations (versus future productivity) in various fields outside of teaching ranged from 0.33 to 0.40 across multiple studies; Glazerman et al. (2010) also cited Goldhaber and Hansen (2010) showing that VAM has produced correlations among teacher evaluation systems "between 0.20 and 0.60 across multiple studies, with most estimates lying between 0.30 and 0.40" (p. 8). Glazerman et al. (2010) justified the use of VAM within teacher evaluation systems regardless of lower reliability results because outcomes have been comparable to other professions and were used to make high-stakes decisions in those professions.

However, other researchers have pointed out that measuring the outcomes of teaching has been prone to errors that introduced noisy statistical features compared to other professions and reliability stemming from teacher evaluation systems, and VAM has been volatile as a result (Haertel, 2013; McCaffrey et al., 2009; Rubin, Stuart, & Zanutto, 2003). In terms of lower reliability measurements, teachers have been less apt to

buy into teacher evaluation systems if their ratings jump unexpectedly from year to year (Amrein-Beardsley & Collins, 2012; Darling-Hammond, Amrein-Beardsley et al., 2012).

Researchers have expressed concern that although some instability should be expected based on a teacher's year-to-year performance (Sass, 2008), lower reliability measurements have caused concern for teacher evaluation systems using VAM (Haertel, 2013; McCaffrey et al., 2009). If teacher evaluation systems have not been able to reliably inform teachers of what teaching practices and teacher leadership are successful and which ones are not, there has been a lack in the quality of feedback and improvement (McCaffrey et al., 2009).

When reliability measures have been in question, there have been consequences for the federal, state, and local systems hoping to implement reform measures to improve teaching practices and teacher leadership on the basis of teacher evaluation systems. For instance, McCaffrey et al. (2009) stated,

> Although some year-to-year variability in teacher performance is to be expected, if outcome-based measures of teacher effectiveness are highly variable, their efficacy in high-stakes personnel decisions will be limited. For example, there are proposals to use estimates of teachers' effects on student test scores or value added to determine which teachers are granted tenure and which are dismissed after an initial probationary period. If value-added measures vary over time, a tenure policy based on a short time frame could lead to the dismissal of many truly effective teachers and the retention of others who prove to be relatively ineffective in boosting achievement. Similarly, if variability in outcome-based measures over time leads to wide swings in who is rewarded, teachers will view merit-based pay plans as largely random, greatly reducing any incentive effects of

> pay-for-performance systems. Moreover, the source of the year-to-year variability (e.g., true variation in teacher performance versus sampling error in the annual measures) will have implications for how best to use the effectiveness measures for evaluating teachers. (p. 573)

Amrein-Beardsley and Collins (2012) carried out a mixed-methods investigation in HISD specifically analyzing data surrounding the dismissal of four teachers and the teacher evaluation system in place at the time (which used EVAAS®). Amrein-Beardsley and Collins discovered instability of EVAAS® outcomes within the HISD teacher evaluation system.

Building on Amrein-Beardsley and Collins (2012), Collins (2014) carried out a study involving the HISD teacher evaluation system and perceptions regarding various aspects of it including its effects on teaching practices and teacher leadership. Collins's surveyed sample included N=874 and included questions about the reliability of the HISD teacher evaluation system and EVAAS®**.** Teachers in the HISD expressed discontent with reliability outcomes of the HISD teacher evaluation system because EVAAS® scores were unstable (Collins, 2014).

Collins (2014) found that almost half (46.2%) of the HISD teacher sample perceived their EVAAS® outcomes as inconsistent and cited changing teaching assignments and student characteristics (especially ELL) as reasons for the unreliable outcomes of EVAAS® concurrent with the HISD teacher evaluation system. Teachers in the HISD relayed concerns about not understanding why EVAAS® outcomes would change unexpectedly while their teaching practices and teacher leadership roles were consistent year to year (Collins, 2014). One teacher summed up the confusion and

negative effects on teaching practices and teacher leadership caused by reliability issues and receiving merit pay based on the HISD teacher evaluation system and EVAAS® outcomes:

> I do what I do every year. I teach the way I teach every year. [My] first year got me pats on the back. [My] second year got me kicked in the backside. And for year three my scores were off the charts. I got a huge bonus, and now I am in the top quartile of all the English teachers. What did I do differently? I have no clue. (Amrein-Beardsley & Collins, 2012, p. 15)

The qualitative input of teachers in the HISD displayed the negative impacts of a perceived low reliability within teacher evaluation systems that used EVAAS® (Amrein-Beardsley & Collins, 2012; Collins, 2014). Teachers were confused about why their outcomes changed, and Amrein-Beardsley and Collins (2012) likened the statistical reliability of EVAAS® as "roughly the same as the flip of a coin" (p. 15) for year-to-year classification of teachers. Amrein-Beardsley and Collins also cited several studies showing that after 3 years of data there was still a "25% risk of misclassification" (p. 15). With this confusion, it was difficult for teachers to be confident that their teaching practices and teacher leadership roles were the drivers behind EVAAS® outcomes (Amrein-Beardsley & Collins, 2012).

It has been recommended in the literature that a 3-year average of VAM outcomes be used within teacher evaluation systems in order to supply teachers with more reliable and credible data pertaining to their performance, rather than year-to-year outcomes which tend to be unreliable (Kane et al., 2010; McCaffrey et al., 2009; MET, 2010). However, after an in-depth study in Tennessee (which will be elaborated upon in

following sections), Xu (2000) made the following comments about such an approach:

> The poor reliability of the value-added scores is obvious. While TCAP test results have been relatively stable over the years, value-added scores swing wildly. The 1994 and 1995 achievement level scores were highly correlated, both at the district level and at the school level (r is above 0.9). One-year gains from 1994 to 1995 were hot highly correlated. The correlation coefficients for math gains were less than 0.1 in the elementary school sample, and less than 0.15 in the district sample. One-year math gains had correlation coefficients of 0.23 among the districts and 0.22 among the elementary schools. Extreme examples were found in some smaller schools in which single-year value-added scores were negative in one year and over 200 percent in the year before and after. Three-year averaging of value-added scores reduces the year-to-year variation, but does not guarantee better quality of the value-added scores. The average of two or three bad guesses is not necessarily a better guess, unless the guesses are unbiased. The way value-added scores are calculated dictates that the result is not going to be very stable. (pp. 89-90)

Xu's (2000) insight has made it important to analyze the perceptions of teachers in order to investigate their perceptions of any possible instability in VAM outcomes and how teaching practices and teacher leadership have been affected as a result. The collection of qualitative data from teacher perceptions pertaining to teacher evaluation systems incorporating VAM has been missing and Amrein-Beardsley and Collins (2012) and Collins (2014) attempted to begin collecting teacher input. Quantitatively, there have been plentiful data collected and analyzed showing statistical reliabilities of teacher evaluation systems using VAM and the effects on teaching practices and teacher

leadership.

Haertel (2013) pointed out that measuring the correlation of VAM outcomes from two points in time (from year to year) has been a statistical methodology employed to measure reliability of VAM within teacher evaluation systems. In doing so, researchers have uncovered similar findings (Ballou, 2005; Koedel & Betts, 2007; McCaffrey et al., 2009). In most studies that follow, researchers divided teachers into groups of quartiles or quintiles (or rarely deciles) and analyzed movement in teacher rankings among groups in order to analyze reliability measurements. In conjunction with this approach, authors also provided the movement of teachers among rankings and identified the effects on teaching practices and teacher leadership.

For instance, Ballou (2005) analyzed the accuracy of teacher rankings between quartiles from TVAAS® outcomes using data from a moderately large school district in Tennessee throughout the 1998 and 1999 school years. Ballou provided the graph in Appendix M1 displaying the percentage of teachers in reading and mathematics who remained in the bottom (lightly shaded) and top quartiles (darkly shaded) between the 1998 and 1999 school year.

Ballou (2005) specifically found that almost 40% of reading teachers and about 40% of mathematics teachers remained in the first quartile in 1999 based on 1998 data, while a little more than 60% of reading teachers and about 60% of mathematics teachers were dispersed among the second through fourth quartiles. Movement in reading and mathematics from the top quartile down to the remaining first through third quartiles in 1999 (based on 1998 data) was similar but somewhat more stable (Ballou, 2005).

The findings in Ballou (2005) were important because they showed statistical reliability of VAM outcomes within teacher evaluation systems and some effects on

teaching practices and teacher leadership. Ballou concluded that although the reliability of TVAAS® could be considered questionable, it was unrealistic to expect TVAAS® as used within teacher evaluation systems would be perfect. Ballou also asserted that other components of teacher evaluation systems were not historically perfect and perhaps VAM offered a more reliable system than historical approaches to teacher evaluation. The effects of unreliable data on teaching practices and teacher leadership could cause confusion about the performance of teachers (Ballou, 2005).

Ballou (2005) cautioned using TVAAS® outcomes within teacher evaluation systems for high-stakes purposes and advised TVAAS® be used within a comprehensive approach using multiple measures. Ballou also mentioned that given the reliability measures of TVAAS® outcomes within teacher evaluation systems, teaching practices and teacher leadership could be negatively impacted by resistance from teachers and an inclination from teachers to "game the system" (p. 23). Ballou showed that a lack of reliability in VAM outcomes within teacher evaluation systems could hinder credibility among teachers.

Koedel and Betts (2007) and Aaronson, Barrow, and Sander (2007) both generated reliability data for teacher evaluation systems using VAM data similar to Ballou (2005); and both studies offered insights into the effects of reliability on teaching practices and teacher leadership. Aaronson et al.'s analysis relied on a sample size of 589 math teachers in CPS and used data from eighth- and ninth-grade students between the 1996-1997 and 1998-1999 school years while Koedel and Betts (2007) analyzed data for 941 teachers in both mathematics and reading between the 2000-2001 and 2001-2002 school years. The data from Koedel and Betts (2007) and Aaronson et al. are listed in

Appendix N1 and Appendix O1, respectively, and shared similarities in their approach.

In both Koedel and Betts (2007) and Aaronson et al. (2007), if reliability measures were perfectly stable, it would be expected that the added diagonals among years would "equal 100 percent and all off-diagonal entries would all equal 0" (Koedel & Betts, 2007, p. 30). Authors of both studies observed this was not the case. However, authors of both studies attributed some of the instability in teacher rankings to teacher turnover.

Koedel and Betts's (2007) table in Appendix N1 was similar to Aaronson et al.'s (2007) in Appendix O1 with the latter being more stable. Koedel and Betts (2007) made a unique insight that applied to both their data and Aaronson et al.'s–the groupings with the highest stability occurred at the northwestern and southeastern locations on the table where the teachers with the lowest and highest VAM outcomes were located, and these groups of teachers would be most likely to retain their position as a result.

Koedel and Betts (2007) remarked that although their reliability measures looked "bleak" (p. 31), those teachers with the highest and lowest quality VAM outcomes could be "targeted" (p. 31) by teacher evaluation systems in order to remediate teachers with the lowest quality of teaching practices and teacher leadership while learning from those with the highest. Although in analyzing their data, Aaronson et al. (2007) did not make mention of this application, the same conclusion would apply.

McCaffrey et al. (2009) carried out a significant study in which they measured the reliability of VAM within teacher evaluation systems in five counties in Florida. Specifically, McCaffrey et al. (2009) investigated the effects of different models and different student assessments on the reliability of VAM outcomes alongside analyzing the inter-temporal variation of teacher effect estimates and the stability of VAM teacher

ratings over the course of 4 years (rather than 2 years as with Koedel and Betts [2007] and Aaronson et al. [2007]). The data pertaining to year-to-year inter-temporal variation from McCaffrey et al. (2009) are supplied in Appendix P1 for review.

The advantage of McCaffrey et al.'s (2009) study was the multiple models testing VAM outcomes that included covariates adjusting for student characteristics and their complete versus partial persistence over time. McCaffrey et al. (2009) noted the range of pooled correlations across 4 years was mostly between 0.2 and 0.5 for elementary school teachers and between 0.3 and 0.6 for middle school teachers, which displayed moderate correlation results. McCaffrey et al. (2009) also discovered that the model employing student covariates with partial persistence yielded the highest correlations (because of smaller sampling errors) for all counties–about 0.10 to 0.14 higher in elementary and about 0.23 higher in middle school.

McCaffrey et al.'s (2009) outcome was significant because policymakers and education officials have had to choose between the most sensible VAM formulas and applications to use within teacher evaluation systems in order to improve teaching practices and teacher leadership (Haertel, 2013; McCaffrey et al., 2009). Stakeholders have demanded outcomes that were reliable (Haertel, 2013). Also, McCaffrey et al. (2009) found that granting teachers tenure based VAM outcomes would improve teaching practices and teacher leadership, and student achievement would increase by about 0.04 of a standard deviation unit of test scores.

Haertel (2013) recognized that for a reliability measurement of 0.5 for 2 or more successive years, half the variation could be attributed to teacher performance; while the other half could be attributed to noise stemming from unmeasured variables. According to Haertel, this outcome would have significantly negative implications for using teacher

evaluation systems and VAM to improve teaching practices and teacher leadership. Specifically, when teacher evaluation systems have lacked the ability to provide teachers reliable feedback, the quality of feedback has the come into question (Haertel, 2013). This has been a general theme detected among reliability studies in the literature, and it has been a continuous line of reasoning by authors that the validity and quality of feedback from teacher evaluation systems using VAM have depended on the reliability of outcomes (Aaronson et al., 2007; Koedel & Betts, 2007; McCaffrey et al., 2009).

Haertel (2013) elaborated upon McCaffrey et al.'s (2009) research by building bar graphs that translated McCaffrey et al.'s (2009) year-to-year correlation measures into the stability of teacher rankings in quintiles. Haertel's analysis is placed in Appendix Q1 for reference.

McCaffrey et al. (2009) and Haertel (2013) dually noted that one-third of the teachers in the bottom quintile stayed, while about 10% of teachers originally in the bottom quintile moved to the top quintile (top graph); likewise, on-third of the teachers in the highest quintile stayed, while about 10% of teachers originally in the top quintile moved to the bottom quintile (bottom graph). Because of the instability in teacher ratings, both McCaffrey et al. (2009) and Haertel notated the negative effects lower reliability and stability measures bore on teacher evaluation systems using VAM. Most notable, McCaffrey et al. (2009) and Haertel found that without the ability to reliably measure teacher performance, teachers would not be able to improve teaching practices and teacher leadership without trustworthy baseline data.

Teacher evaluation systems that have used VAM and were prone to swings in the rankings of teachers have had a diminished ability to help those teachers who need it most and reward those teachers who deserve recognition (Haertel, 2013; McCaffrey et

al., 2009). As a result, many researchers have warned against using VAM measures alone to make high-stakes decisions (Haertel, 2013; Koedel & Betts, 2007; McCaffrey et al., 2009; Sass, 2008).

Ballou, Mokher, and Cavalluzzo (2012) investigated the impact of using different VAM formulas and applications using a large sample of teachers (N=2,221) and students (N=519,468) in order to compare a VAM "true" model that included full specification of covariates to a VAM "misspecified" model that included limited covariates. This approach was parallel to McCaffrey et al. (2009).

Ballou et al. (2012) specifically investigated the effects of both models of VAM on teachers at each end of the performance distribution and found that of the 221 teachers in the top 10% using the "true" model, 92 (42%) were not classified as such in the "misspecified" model. Likewise, 88 teachers classified in the bottom 10% by the "true" model were no longer classified as such in the "misspecified" model (Ballou et al., 2012). With these reliability results, Ballou et al. (2012) cautioned implementing teacher evaluation systems targeting high and low performers using VAM techniques to distinguish among teachers based on performance.

Teacher evaluation systems have failed to improve teaching practices and teacher leadership when VAM outcomes relayed unpredictable outcomes, and Ballou et al.'s (2012) study was disconcerting because the authors showed that different models of VAM misclassified bottom and top performers. The bottom and top performers have been the groups most targeted by teacher evaluation systems for actions to either improve or reward performance (Ballou et al., 2012; Koedel & Betts, 2007). Furthermore, merit pay programs or devised sanctions for chronic low performers could falter when relying upon unreliable outcomes provided by teacher evaluation systems using VAM (Ballou et

al., 2012).

In comparing the reliability of VAM across different standardized tests, there have been mixed results. Gill, Bruch, and Booker (2013) and MET (2010) found that the choice of standardized tests produced similar VAM outcomes that were correlated. Among other researchers, the reliability of VAM has shown to swing among teacher ratings when different standardized tests have been used to calculate VAM outcomes within teacher evaluation systems (McCaffrey et al., 2009; Papay, 2011; Sass, 2008; Stuit, Berends, Austin, & Gerdeman, 2014). This concern may be alleviated in the future by improving the quality of standardized tests used to calculate VAM within teacher evaluation systems (Haertel, 2013). Both the Partnership for Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (SBAC) have functions to improve standardized tests centering on content from the Common Core; and as the quality of standardized tests increases, their reliability within VAM estimates should also (Haertel, 2013).

Sass (2008) found that using different tests created instability among teacher rankings within quintiles using the low-stakes Stanford Achievement Test (NRT) versus the high-stakes state-mandated standardized tests derived from the Sunshine State Standards (SSS). Sass data for Hillsborough County in Florida is placed in Appendix R1 for reference.

When comparing the NRT and the SSS outcomes and ranking of teachers, the data found 70% of teachers ranked in the top two quintiles, 69% ranked in the bottom two quintiles, and the remaining was spread among the other quartiles. Sass (2008) calculated a correlation of 0.48 between the outcomes of the two tests, which resulted in a moderate reliability level. Although Sass concluded that “different tests result in different

teacher rankings" (p. 5) in terms of the reliability, Sass also stated that the "relative instability" (p. 5) in teacher rankings does not necessarily eliminate a role for VAM within teacher evaluation systems and merit pay programs. Averaging data among 3 years' worth of data has shown to add stability to VAM outcomes and has given teacher evaluation systems the ability to identify bottom and top performers for the sake of remediation or rewards (McCaffrey et al., 2009; Sass, 2008).

Sass (2008) also found that the correlation of VAM to NRT was significantly lower (0.27) than that of the comparing the overall correlation between NRT and SSS (0.48). Sass hypothesized that the stakes attached to the tests within teacher evaluation systems (SSS) may be one explanation and that such stakes may cause teachers to modify their teaching practices and teacher leadership to concentrate more on fulfilling the outcomes of high-stakes tests versus low-stakes tests.

MET (2010) carried out a significant investigation studying the effects of different standardized tests on VAM outcomes within teacher evaluation systems. MET (2010) also detailed how the reliability of VAM used within teacher evaluation systems could be used to improve teaching practices and teacher leadership and increase achievement among students.

What made MET (2010) significant was its sample size (N=3,000 teachers); its attention to randomly assigning students to teachers; and its ability to test reliabilities for particular standardized tests, various VAM formulas, and student survey results across different sections and different years. The results of MET (2010) are assembled in Appendix S1 in a data table where the unadjusted "total variance" refer to a persistent (teacher quality effects) component and the nonpersistent (sources of variation) component alongside the "implied variance of stable component" refer to isolated teacher

quality effects across sections or years.

MET (2010) advised that when between-section or between-year correlation outcomes for VAM were less than 0.5 when comparing the various standardized tests, more than half of the variation was due to "transitory effects rather than stable differences between teachers" (p. 19); and MET (2010) also recognized that their calculated correlations between sections and the prior year were less than 0.5. Regardless of the lower reliability results among the between-section and between-year measurements, MET (2010) concluded that because of the large total variance and assuming a bell-shaped distribution of teacher effects, top quartile teachers would increase average student achievement by 0.18 standard deviations compared to the median teacher.

The total variation and by extension the stable variance was large enough in MET (2010) results to give discriminatory power in judging teachers based on their performance, which was an outcome parallel to Ballou's (2005) conclusion alluded to earlier in this section. Taken together, MET (2010) found that the difference between a low and high quartile teacher was 0.36 standard deviations (0.18 below + 0.18 above), which would be an amount of student achievement gain analogous to closing more than one-third of the Black-White achievement gap in fourth and eighth grade on the NAEP test.

A limitation to this conclusion was that the ability to identify a "top quartile teacher" (MET, 2010, p. 19) consistently from year to year (or section to section) could be difficult if the teacher effects were not stable across tests, sections, or years. Ballou (2005), Koedel and Betts (2007), Aaronson et al. (2007), Sass (2008), and McCaffrey et al. (2009) found movement in the ranking of teachers among quartiles (or quintiles) when assessing the stability of teacher rankings based on VAM outcomes across different

years, VAM formulas, and standardized tests.

Although MET (2010) did include VAM correlations for the four standardized tests between sections and the prior year, data was not found that calculated movement of teachers based on their ratings among quartiles or quintiles. McCaffrey et al. (2009) did provide some insight into the expectation of quartile movement among teachers based on reliability data using calculations extracted from Lockwood et al. (2002):

> If the data are normally distributed and our model holds, the stability coefficient determines the proportion of teachers who will switch quintile ranks across years and how much error there is in predictions of future teacher effects. For example, if the estimated effects have a stability of 0.3, we can expect that about 24, 19, 15, and 10 percent of teachers ranked in the first quintile in one year will be ranked in the second, third, fourth, and fifth quintiles, respectively, in the next year. For an estimated effect with a stability of 0.7 the percentages are 26, 12, 5, and 1. (p. 580)

MET (2010) collected and displayed data showing the section-to-section correlations serving as the reliability coefficient for all four standardized tests and the Tripod student perception survey across sections (Appendix T1).

MET (2010) noted the higher reliability of the student surveys summed at 0.668 across sections. The Tripod student survey correlations were significant because they showed that students were able to identify effective teachers on the aggregate, and their level of reliability was higher than any other indicator in MET (2010). However, given McCaffrey et al.'s (2009) aforementioned calculations, if the teachers in MET (2010) were divided into quintiles based on effectiveness estimates from VAM at a correlation of 0.30 (similar to MET Project outcomes), there would be 32% who would retain their

quintile in the first quartile the following year; while 24% of teachers would move into quintile 2, 19% into quintile 3, 15% into quintile 4, and 10% into quintile 5.

This estimate would make MET's (2010) goal of consistently identifying "a top quartile teacher" (p. 19) more complicated if there was significant movement among quartiles based on teacher rankings. And based on observations from multiple authors, instability in quartile ratings has brought mistrust of teacher evaluation systems and hindered the quality and trustworthiness of feedback (Amrein-Beardsley & Collins, 2012; Collins, 2014; Haertel, 2013).

In analyzing MET (2010) reliability levels, Haertel (2013) made the following relevant comments:

> On standardized tests with stakes for students, reliability coefficients of at least .80, preferably .85 or .90, are the goal. A coefficient of .80 means that 80% of the variation in scores is attributable to real differences in the attribute the test measures, and only 20% is measurement error. Value-added reliabilities of .2 to .5 imply that as little as 20% of the score variation is attributable to the quality the scores are measuring and as much as 80% is due to measurement error.
>
> Of course, reliability of VAM scores can be increased considerably by pooling results over 2 or 3 years. If the reliability of single year VAM scores were .30, say, then a 2 year rolling average should have a reliability of roughly .46, and a 3 year rolling average, roughly .56—these numbers are still not very good, but they are much improved over single year estimates. Unfortunately, many VAM implementations have relied on results from just a year at a time. (p. 19)

While the importance of reliability in the quality and credibility of feedback offered to teachers has been an important topic, the parameters of measurements have

seemingly been agreed upon (Glazerman et al., 2010; McCaffrey et al., 2009; MET, 2010). The conditions of using teacher evaluation systems using VAM have centered upon whether it has been philosophically acceptable compared to how reliability estimates have been used in professions outside of teaching (Glazerman et al., 2010).

Significant studies have relied upon the extension of this logic in justifying the use of VAM within teacher evaluation systems (McCaffrey et al., 2009; MET 2010). MET (2010) cited Smith and Schall (2000) to compare the between-season correlations for batting averages of Major League baseball players (0.36) and earned-run averages (ERA) for pitchers (0.31) which were lower than MET's (2010) correlations for mathematics and higher than the English tests. McCaffrey et al. (2009) did the same citing Hoffman, Jacobs, and Gerras (1992) showing the correlation for between-season batting averages in the range of 0.32-0.48 and for between-season ERAs for pitchers in the range of 0.12-0.45. McCaffrey et al. (2009) pointed out that some professions require "repetitive tasks over short periods of time" (p. 593), which produced higher correlations at least in the short term.

In discussing the instability of reliability outcomes in teacher evaluation systems that use VAM, the literature has concentrated on attempting to measure the amount of statistical error and how error levels have affected teaching practices and teacher leadership (Amrein-Beardsley & Collins, 2012; Ballou, 2005; Collins, 2014; Haertel, 2013; McCaffrey et al., 2009). One such study was the *Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains,* a federally funded study carried out by Mathematica Policy Research; and it measured the amount of error teacher evaluation systems could expect to overcome in measuring VAM (Schochet et al., 2010).

The Schochet et al. (2010) investigation was relevant and significant for three reasons. First, Schochet et al. included simulated error rates for EVAAS®, which has been the VAM model used within the county of the primary research of this study. Second, Schochet et al. concentrated on determining error rates for identifying high performers and low performers. Last, Schochet et al. notated the level of statistical error using VAM within teacher evaluation systems and the effects on teaching practices and teacher leadership.

Specifically, Schochet et al. (2010) found when using typical data involved with VAM measurements, there was an error rate of about 26% when 3 years of data were used. Schochet et al. made the following observation applying Type I errors as false positives and Type II errors as false negatives:

> Type I and II error rates for teacher-level analyses will be about 26 percent if three years of data are used for estimation. This means that in a typical performance measurement system, more than 1 in 4 teachers who are truly average in performance will be erroneously identified for special treatment, and more than 1 in 4 teachers who differ from average performance by 3 months of student learning in math or 4 months in reading will be overlooked. In addition, Type I and II error rates will likely decrease by only about one half (from 26 to 12 percent) using 10 years of data. (p. 35)

Schochet et al. (2010) also identified the negative effects of error rates on the stability of teacher ratings within teacher evaluation systems employing VAM. When error rates were large, teaching practices and teacher leadership suffered from inconsistent feedback and could "impose undue psychic and economic costs on teachers who are falsely identified as low performers" (Schochet et al., 2010). This could

discourage future teachers with high-quality teaching practices and teacher leadership from entering the teaching profession (Schochet et al., 2010). Also, the existing pool of teachers could endure reduced morale when ratings are unstable due to a high number of errors from teacher evaluation systems using VAM (Schochet et al., 2010).

In summary, to validate subjective ratings from principals (Peterson, 2000), VAM has been an indicator of teacher effectiveness put into place using objective outcomes centered on student learning (Goe, 2008). The goal has been to minimize the inflation of principal ratings that have historically been shown to plague subjective teacher ratings from principals in traditional teacher evaluation systems (Weisberg et al., 2009).

VAM has shown to successfully differentiate among teachers (Bock et al., 1996; Daley & Kim, 2010; Sanders & Rivers, 1996; TNDOE, 2012) in newly revamped teacher evaluation systems parallel to newly developed rubrics shown in the previous section of this literature review dedicated to the quality of classroom observations and evaluation (Batten et al., 2012; Curtis, 2011; Daley & Kim, 2010; Danielson, 2007; Lynn et al., 2013). There has now been a marriage between the ability of principals to use newly designed performance-based rubrics to subjectively differentiate among teachers and the ability of VAM to produce objective estimates of teacher effectiveness stemming from student outcomes. The outcomes of principal ratings and VAM outcomes within teacher evaluation systems have shown weak to moderate correlation levels using statistical tests (Curtis, 2011; Headden, 2011; Holtzapple, 2003; Jacob & Lefgren, 2008; Kane et al., 2010; Milanowski et al., 2004; Sartain et al., 2011; Stronge & Associates, 2013).

The results have harvested positive results in Tennessee where the state has enjoyed historic gains in proficiency and growth outcomes on TCAP (TNDOE, 2012). Tennessee education officials have largely credited the new teacher evaluation systems

where this “marriage” of principal ratings and TVAAS outcomes has improved teaching practices and teacher leadership (TNDOE, 2012).

There has been some debate in the literature regarding the validity of using VAM within teacher evaluation systems on the basis of evidence showing correlation with students’ SES (Goldhaber & Theobald, 2013; O’Donnell, 2013; WCPSS, 2009; Xu, 2000). It was noted that EVAAS® has shown some level of correlation at the teacher and school levels with students’ SES, but this evidence has been contrasted by scatterplots from the SAS Institute showing no correlation. Ongoing research will be needed in order to validate the use of EVAAS® within teacher evaluation systems given evidence of the correlation.

In contrast to evidence of the correlation between VAM and students’ SES, it was shown that teachers in high-poverty schools were as likely to show high growth in teacher evaluation systems using VAM as they would be in low-poverty schools (Sass et al., 2010). This outcome reinforced equity among teachers across the spectrum of a diverse student SES range where any teacher has the potential to equally improve teaching practices and teacher leadership and improve student learning.

In terms of teacher evaluation systems employing VAM and how reliability outcomes have affected teaching practices and teacher leadership alongside student achievement, there were several general themes agreed upon among authors within this section. First, within-year VAM measurements have displayed significant distribution among teachers based on teaching practices and teacher leadership enabling teacher evaluation systems to identify bottom and top performers (Aaronson et al., 2007; Ballou, 2005; Koedel & Betts, 2007; McCaffrey et al., 2009; MET, 2010). This has enabled teacher evaluation systems an opportunity to improve teaching practices and teacher

leadership for bottom performers and reward and learn from top performers using VAM as a driver.

Second, teacher evaluation systems using VAM for between-year measurements have produced enough instability in teacher rankings that authors have suggested a cautious approach in using VAM singularly to evaluate teaching practices and teacher leadership for teachers, especially for high-stakes personnel decisions (McCaffrey et al., 2009; MET, 2010; Sass, 2008; Schochet et al., 2010). These authors have also recommended that to reduce measurement error, VAM outcomes should be used in conjunction with multiple measures including administrator judgments and student survey data (MET, 2010; Schochet et al., 2010). Some researchers caution the use of VAM within teacher evaluation systems for low-stakes decisions on the basis of low reliability and for the risk of causing unintended consequences (Amrein-Beardsley & Collins, 2012; Collins, 2014; Haertel, 2013).

Lastly, when reliable feedback has been offered to teachers on the basis of reliable outcomes from teacher evaluation systems using VAM, there has been an improvement in teaching practices and teacher leadership (Koedel & Betts, 2007; McCaffrey et al., 2009), and the opposite effect and outcome has also been shown (Sass, 2008). Along the same line of reasoning, reliable VAM outcomes within teacher evaluation systems have built confidence and validity in associated processes such as administrator feedback given to teachers and teacher buy-in (McCaffrey et al., 2009).

Unintended consequences of teacher evaluation systems incorporating VAM. Alongside issues involving the validity and reliability of teacher evaluation systems using VAM and the effects on teaching practices and teacher leadership, a predominate subject in the literature has been unintended consequences caused by VAM in teacher evaluation

systems and the resulting effects on teaching practices and teacher leadership (Amrein-Beardsley & Collins, 2012; Baker et al., 2010; Collins, 2014). The unintended consequences have been difficult to measure in the literature because they have been predictive in nature, but there has been some initial qualitative evidence in this realm collected by researchers (Amrein-Beardsley & Collins, 2012; Collins, 2014; Darling-Hammond et al., 2011). Future research will be needed on this topic in order to detect unintended consequences of VAM in teacher evaluation systems and to what extent these consequences have affected teaching practices and teacher leadership (Amrein-Beardsley & Collins, 2012; Baker et al., 2010; Collins, 2014).

After surveying the literature, there have been at least five categories in which future potential unintended consequences could fall, and each purportedly could have negative effects on teaching practices and teacher leadership at the present time and in the future. There has been an umbrella theme that these unintended consequences could be the result of a competitive element recently interjected within teacher evaluation systems using VAM (Amrein-Beardsley & Collins, 2012; Baker et al., 2010; Collins, 2014).

As Murphy (2012) observed in relating "attainment-based accountability" to student proficiency levels, recently revamped teacher evaluation systems using VAM have relied on a competitive element to improve teaching practices and teacher leadership:

> Because attainment-based accountability systems compare different populations and do not control for prior achievement, they are a poor measure of the effect of an individual teacher's contribution to student achievement. By contrast, VAM is designed to evaluate the learning growth of a population of students, control for factors beyond the teacher's control, and isolate the contributions that teachers

> make toward student test score gains. As a result, VAM provides a measure that enables comparisons across different teachers, and inferences can be drawn such as teacher X stimulates more learning than teacher Y. (p. 6)

Scherrer (2012) also noted the following relative to the competitive element interjected within teacher evaluation systems using VAM:

> Not only does VAM fail to eliminate some of the concerns associated with high-stakes testing, VAM introduces some new concerns. VAM distributes teachers onto a normative scale. That means teachers are compared to each other and not some criteria. This type of distribution accords a limited number of above-average slots. Stated differently, no matter how much teachers improve, half of them will always be below average. This creates a highly competitive situation. Proponents of high-stakes accountability will assert this type of highly competitive environment is exactly what will produce results. One must, however, consider the likely consequences of setting teachers against each other. (pp. 59-60)

First, there has been a concern that teacher evaluation systems using VAM could be biased favoring teachers with more affluent students (Ladd & Walsh, 2002; McCaffrey, 2013) as was the case with accountability measures wrought by the NCLB Act of 2002 (Braun, 2005). As a result, teachers may be attracted to working in wealthier schools (Feng, Figlio, & Sass, 2010; Ladd et al., 2002; McCaffrey, 2013) and may attend to specific students who could yield the most growth (Baker et al., 2010; Kuppermintz, 2003). Consequently, some authors worry that the opposite may be true in the future–that a "ceiling effect" may make it difficult for teachers to show growth among more affluent students who already achieve high proficiency scores (Baker et al., 2010; McCaffrey, 2013).

In reviewing this literature, there has been a lack of quality evidence affirming the transfer of teachers from schools with high FRL rates to schools with lower rates or vice versa, but future research will be needed. Sass et al. (2010) and Glazerman et al. (2013) showed evidence that teachers with high VAM outcomes in low-poverty schools have been as likely to produce the same results in a high-poverty school. Although high poverty schools have higher teacher attrition (Smith & Ingersoll, 2004), there has been no evidence in the past that teacher evaluation systems using VAM have been responsible for the migration or attrition of teachers. Also, there has been no evidence uncovered in the literature that teachers are targeting specific types of students for the sake of producing high growth scores, but future research will be necessary in this area.

Second, authors have expressed concern that because of competition among teachers and schools, school personnel may use inappropriate or unethical methods to create higher growth outcomes among students (GAO, 2013; Nichols, Glass, & Berliner, 2005; Scherrer, 2012). There has been cheating on state-mandated standardized tests uncovered recently (Bowers, Wilson, & Hyde, 2011; GAO, 2013; Gillum & Bello, 2011) and in the past (Nichols et al., 2005) as a result of teachers and administrators gaming the system in order to inflate outcomes within teacher evaluation systems.

A recent investigation (Bowers et al., 2011) in the Atlanta Public Schools (APS) uncovered systemic pressures to increase scores on state standardized tests and found a combined 82 teachers and principals who confessed to cheating in schools throughout the Atlanta school district during the 2008-2009 school year (GAO, 2013). Bowers et al. (2011) highlighted a “culture of fear” (p. 356) that resided among teachers and principals in APS that led to either participation in altering student answer sheets on state standardized tests or ignoring that the process was occurring.

GAO (2013) surveyed testing officials in state educational agencies across the country and found that for the previous 2 years before the study, 40 states reported allegations of cheating, while 33 confirmed at least one case of cheating. Moreover, 32 states reported canceling test scores in confined schools as a result of suspected or confirmed cases of cheating within this time period (GAO, 2013).

When teachers and administrators have cheated in the past, the consequences of their actions negatively affected teachers, teaching practices and teacher leadership, and the teaching profession (Nichols et al., 2005). The "culture of fear" (Bowers et al., 2011, p. 356) apparent in the APS system placed teachers and administrators in awkward positions to produce outcomes on state standardized tests "by whatever means necessary" (Bowers et al., 2011, p. 356) and regardless of any measure of teaching practices and teacher leadership. Headden (2011) outlined the dangers of score inflation given newly designed teacher evaluation systems that use VAM, one of which was that teachers who inherited students with inflated growth scores would be expected to produce an inflated trajectory of growth placing such a teacher at a disadvantage.

One characteristic of the teachers and administrators in the APS system was they viewed the state standardized tests on which they cheated as an unimportant measure while tending to what they considered important to meet the needs of students in their classrooms and schools (Bowers et al., 2011). In investigating cheating cases, Nichols et al. (2005) noted that it was "plausible that teachers and administrators are trying to resist a system they see as corrupt and unfair" (p. 24). Specifically, Aviv (2014) recounted narrative from an attorney who sat in on the interviews of suspected teachers and principals within the APS system:

Righton Johnson, a lawyer with Balch & Bingham who sat in on interviews, told

> me that it became clear that most teachers thought they were committing a victimless crime. "They didn't see the value in the test, so they didn't see that they were devaluing the kids by cheating," she said. Unlike recent cheating scandals at Harvard and at Stuyvesant High School, where privileged students were concerned with their own advancement, those who cheated at Parks were never convinced of the importance of the tests; they viewed the cheating as a door they had to pass through in order to focus on issues that seemed more relevant to their students' lives. (para. 55)

As the stakes of state-mandated standardized tests have recently increased under the RttT initiative, it will require future research to uncover the effects of the pressure associated with high-stakes standardized testing used within teacher evaluation systems for VAM calculations. If county and district education leaders have used data from state-mandated standardized testing as an "abusive and cruel weapon to embarrass and punish" (Bowers et al., 2011, p. 354) teachers and administrators as was the case reported within the investigation of the APS system, how would teachers and administrators respond in the future?

Third, another concern related to the previous has been that a concentrated effort on increasing test scores could potentially narrow the curriculum, resulting in teachers focusing on preparing their students for state-mandated standardized tests used to calculate VAM within teacher evaluation systems (Baker et al., 2010). Teachers have expressed discontent with preparing students for standardized tests and "teaching to the test" (Jones & Egley, 2007), which purportedly has the ability to inflate scores on state-mandated test scores (Baker et al., 2010; Firestone, Monfils, & Schorr, 2004; Shepard, 1990).

Questions in the literature have surrounded the value of preparing students for tests using techniques of rote memorization rather than adopting critical-thinking processes (National Research Council, 1999). As cited by Jones and Egley (2007), the National Research Council (1999) noted, “Tests often reinforce memorizing rather than understanding” (p. 234).

Some unintended consequences on teaching practices under the accountability demands of high-stakes testing were examined during the era of the NCLB Act when state-mandated standardized testing outcomes were used to grade schools rather than teachers (Jones & Egley, 2007; Nichols et al., 2005). While the NCLB Act concentrated on improving outcomes by holding schools accountable, the RttT initiative has sought to improve outcomes by holding teachers accountable for student learning as measured by state-mandated standardized tests.

Because of the connected approach of the NCLB Act and the RttT inititaive, the effects of high-stakes testing on teaching practices and teacher leadership under the RttT initiative could be inferred from historical research (Nichols et al., 2005) involving outcomes from the NCLB Act of 2002 (Braun et al., 2010). High-stakes testing for the sake of teacher evaluation ratings has been amplified under the RttT initiative (Baker et al., 2010). The vantage of this approach may give observers a preview of the future pressures exerted on teachers under the RttT initiative, which was the impetus for newly revamped teacher evaluation systems with VAM. Such an approach may also provide clues about the effects of high-stakes state-mandated standardized tests on teaching practices and teacher leadership in the future under the RttT initiative.

For instance, Jones and Egley (2007) examined survey data from Florida teachers (N=708) in an attempt to describe the effects of the Florida Comprehensive Assessment

Test (FCAT) on teaching practices and teacher leadership within the era of the NCLB Act. The relevant and important conclusions Jones and Egley found was that 96.7% of the sampled teachers felt between "some pressure" and "a lot of pressure" to improve FCAT scores, while 61.8 responded specifically that they felt "a lot of pressure."

Also, Jones and Egley (2007) asked teachers to what extent ("negatively influences me," "does not influence me," "positively influences me") the FCAT influenced their ability to use what teachers thought to be "effective teaching methods" (p. 237). On average, reading teachers reported that the FCAT negatively affected their ability to use effective teaching practices while writing and mathematics teachers reported the FCAT did not affect their ability to choose effective teaching practices.

More importantly, Jones and Egley (2007) conducted *t* tests to measure the amount of time teachers spent teaching to prepare their students for the FCAT in conjunction with the pressure teachers felt to increase FCAT scores. Jones and Egley summarized their findings with the following observations:

> Teachers who felt the most pressure reported spending a significantly higher percentage of their instructional time on test-taking strategies in reading, writing, and mathematics. The percentage of time spent teaching test-taking strategies in reading was 40.9 percent for teachers who felt the most pressure and 28.3 percent for teachers who felt some pressure ($t=6.05$, $p<.001$, $d=0.54$). Similarly, the percentage of time spent teaching test-taking strategies in writing was 45.9 percent for teachers who felt the most pressure and 32.6 percent for teachers who felt some pressure ($t=5.18$, $p<.001$, $d=0.47$). The percentage of time spent teaching test-taking strategies in mathematics was 46.1 percent for teachers who felt the most pressure and 33.6 percent for teachers who felt some pressure

(t=5.61, p<.001, d=0.51). (p. 42)

Data has begun to trickle in regarding unintended consequences of high-stakes testing used within teacher evaluation systems for the sake of calculating VAM outcomes. Oakes and Robertson (2014) surveyed North Carolina teachers (N=800) statewide in order to measure teacher perceptions regarding NCEES and found over 70% of teachers agreed with their ratings in Standards 1-5, while 49% agreed less that Standard 6 EVAAS® ratings accurately reflected the quality of their teaching practices. Relevant to the topic of increasing test scores for Standard 6, Oakes and Robertson found that 75% of teachers agreed they spent too much instructional time preparing students for state-mandated standardized tests; and 85% of the respondents thought students spent too much time taking state-mandated standardized tests. Oakes and Robertson also found that 11% of teachers agreed that students benefited from state standardized tests and that the tests were worth the time and effort.

Firestone et al. (2004) was interested in finding out how the pressure on teachers associated with standardized tests affected their teaching practices and surveyed and interviewed between 250 and 300 teachers over the course of 3 years to measure the effects. The authors found that pressure to improve test scores influenced teachers to elect didactic, traditionally teacher-led instruction; whereas teachers who felt less pressure to improve test scores chose more innovative, inquiry-oriented approaches in pedagogy (Firestone et al., 2004). Another important conclusion of Firestone et al. was the pressure applied by principals and how it affected teachers. Firestone et al. found that with increasing pressure to increase test scores, teachers responded with more long-term (throughout the year) and short-term (before administration) test preparation activities woven into inquiry-oriented teaching practices.

MET (2010) was interested in how value-added outcomes for teachers compared among student assessments, specifically asking whether teachers who showed higher value-added outcomes among their students on state standardized tests were able to show the same on alternative assessments. In mathematics, MET (2010) research involved comparing teacher value-added outcomes on state mathematics tests to value-added outcomes on the Balanced Assessment in Mathematics (BAM), which was designed to test a student's higher-order conceptual understanding. In ELA, the same principle was applied comparing value-added outcomes on state ELA tests to the Stanford 9 Open-Ended Reading test which, like BAM, was designed to test a student's higher-order conceptual understanding.

The rationale behind MET's (2010) approach was that if teachers were overly or exclusively preparing their students for state standardized tests, these teachers would have higher value-added outcomes on those tests but lower value-added outcomes on the conceptual BAM and Stanford 9 tests. MET (2010) results showed that this was not the case as those teachers who had higher value-added outcomes on state standardized tests showed higher value-added outcomes on the alternate tests.

Specifically, the persistent effects of teachers between the state mathematics test and the BAM test resulted in a correlation of 0.54 and 0.59 for the state ELA test and the Stanford 9 test (in ELA the results excluded NYC due to changes in state tests) (MET, 2010). The MET (2010) team concluded that the "moderately large" correlation between state tests and alternate conceptual tests indicated teachers were not "teaching to the test" (p. 21) and that students of teachers with higher value-added gains on state tests had gained a better conceptual understanding of the content.

From the aforementioned studies (Firestone et al., 2004; Jones & Egley, 2007;

Oakes & Robertson, 2014), it appeared teachers noted pressure to increase test scores of their students, but it has been unclear whether they chose the best forms of teaching practices and teacher leadership to do so. While MET (2010) showed that increases on state standardized tests do not necessarily mean that teachers are teaching to the test, Jones and Egley (2007) argued "test-based accountability policies should focus less on pressuring educators into compliance and more on providing support through quality professional development" (p. 234), but future research will be necessary to measure the effects of any associated pressure to increase test scores on teaching practices and teacher leadership. It has also been noteworthy that during MET (2010), the related standardized assessments were not meant for high-stakes outcomes, meaning teachers were under no pressure to increase standardized test scores for the good of their own personal evaluations.

Fourth, authors have been concerned that competition among teachers within teacher evaluation systems using VAM has created "disincentives" (Baker et al., 2010, p. 4) for teacher collaboration and has negatively affected teaching practices and teacher leadership as a result. The implementation of PLCs across the U.S. has allowed teachers to significantly improve teaching practices and teacher leadership (Berry, Daughtrey, & Wieder, 2009). Teachers have been shown to rely on other teachers for help more than any other source, and teacher collaboration in PLCs has improved teaching practices and teacher leadership and increased student achievement (Berry et al., 2009; Vescio, Ross, & Adams, 2008).

Although measuring disincentives for teacher collaboration has been in the initial stages, Collins (2014) uncovered teacher perceptions of diminishing teacher collaboration due to the competitive nature of EVAAS® data within the teacher evaluation system

under Collin's study. One respondent in Collins provided his/her perception of the effects of EVAAS® on teacher collaboration within his/her district's teacher evaluation system, which used a pay-for-performance plan in conjunction with EVAAS® outcomes:

> Since the inception of the EVAAS system, teachers have become even more distrustful of each other because they are afraid that someone might steal a good teaching method or materials from them and in turn earn more bonus money. This is not conducive to having a good work environment, and it actually is detrimental to students because teachers are not willing to share ideas or materials that might help increase student learning and achievement. (p. 19)

While Collins (2014) detailed disincentives for teacher collaboration pertaining to using EVAAS® data within a pay-for-performance approach, these concerns have been reflected independent of pay-for-performance approaches (Baker et al., 2010). Even without a financial bonus, EVAAS® could have induced disincentives for teacher collaboration because teachers competitively vie for high student test scores while being at an advantage when students enter their classrooms with lower projected EVAAS® scores calculated based on classes students previously took.

In North Carolina, NCEES was designed to incorporate a school-wide EVAAS® component where 30% of teacher growth scores stemmed from school-wide growth and 70% from their own students. This approach was meant to increase collaboration among teachers and administrators as they worked together to increase school-wide EVAAS® growth.

However, as Lynn et al. (2013) noted, NCDPI recognized when analyzing data that high-performing teachers in low-performing schools were penalized by school-wide

EVAAS® outcomes, while low-performing teachers in high-performing schools received inflated school-wide EVAAS® outcomes. As a result, NCEES no longer incorporated school-wide growth for individual teachers, and a teacher's growth has been completely based on the growth of their own students.

There has been conflicting evidence in Tennessee highlighted in the TNDOE (2012) study compared to that of Collins (2014) and NCDPI noted by Lynn et al. (2013). Unlike Collins, the TNDOE (2012) study recognized and celebrated an increase in teacher collaboration across their newly developed teacher evaluation systems (TEAM, TIGER, COACH, TEM) in conjunction with the use of TVAAS®. Unlike NCDPI, the TNDOE (2012) study specifically found that the use of TVAAS® as a school-wide growth component within teacher personal evaluation scores contributed to historic TCAP growth since the revamping of Tennessee's teacher evaluation systems.

Specifically, the TNDOE (2012) study analyzed qualitative data from principal interviews and noted school-wide TVAAS® "has increased collaboration among teachers and led to a higher emphasis on academic standards in all subjects" (p. 16). Additionally, the TNDOE (2012) study credited TVAAS® and school-wide growth for a "heightened sense of shared responsibility and interdisciplinary collaboration," which conflicted with evidence provided by Collins (2014) and NCDPI as indicated by Lynn et al. (2013).

This has been another area of proposed unintended consequences found in the literature that has produced conflicting evidence; and because of the early stages of newly redesigned teacher evaluation systems, future research will be necessary to describe the effects of incorporating VAM on teacher collaboration. Future research should be grounded in the perceptions of principals and teachers in an attempt to measure the levels of collaboration and its relationship with VAM used within teacher evaluation systems.

The fifth and last of the unintended consequences of using VAM within teacher evaluation systems and the effects of those consequences on teaching practices and teacher leadership lies in what Baker et al. (2010) called teacher "demoralization" (p. 19). Baker et al. admitted this unintended consequence has had limited evidence and has been anecdotal in nature. Amrein-Beardsley and Collins (2012) and Collins (2014) attributed low morale to the use of VAM within teacher evaluation systems when feedback has been confusing to teachers, especially when multiple measures of evaluation did not align in ratings or outcomes (e.g., principal ratings, VAM outcomes, performance pay).

Although teacher job satisfaction was measured at a 20-year low in the recent MetLife Survey of the American Teacher (MetLife, 2012), dropping from 59% who were very satisfied in 2009 to 44% in 2011, and the number of teachers indicating they were "very or fairly likely" (p. 7) to leave the profession increased from 17% in 2009 to 29% in 2011, MetLife provided no specific tie using those statistics to redesigned teacher evaluation systems. More research would be needed on this topic specifically targeting teachers who perhaps have been rated at high levels by their principal but low levels by VAM or vice versa. According to research, a lack of reliability among teacher effectiveness indicators has caused low morale, but there has been a lack of data to make large scale inferences (Amrein-Beardsley & Collins, 2012; Collins, 2014; Nichols et al., 2005).

For unintended consequences of using VAM within teacher evaluation systems and the effects on teaching practices and teacher leadership, Koretz (2008) and Nichols et al. (2005) identified Campbell's Law as a driver. Campbell's Law was developed by. Campbell (1976) who measured the effects of policy decisions within social systems, from which he concluded with the following:

> 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. (p. 49)

Whether teachers have been attracted to certain schools or students; whether cheating or gaming the system occurred; whether teachers have taught to the test; whether there has been diminished collaboration among teachers; or whether there has been low morale and teacher demoralization, Nichols et al. (2005) identified that these unintended consequences have ensued because of an overemphasis on producing quantitative output. Nichols et al. performed a mixed-methodology study showing that the meaning of quantitative data in various social venues became more uncertain as the stakes of producing quantitative data increased.

Research from MET (2010, 2012) has advocated using multiple measures when rating teacher performance within teacher evaluation systems in order to avoid the overreliance on any one indicator and for balance between quantitative (VAM) and qualitative (principal or peer ratings) measures. However, Campbell (1976) recognized the work and input of Ridgway (1956) in casting doubt upon the veracity of multiple measures to alleviate the distortion of unintended consequences within social systems.

For example, principal ratings have been the primary, alternate measure whereby VAM has been compared in correlation and regression studies shown throughout this literature review as with Batten et al. (2012), Batten (2013), Xu (2000), and Lynn et al. (2013). However, Headden (2011) identified that principals in Washington D.C.'s teacher evaluation system were at risk to "adjust scores up or down to compensate for ratings given by master educators" (p. 8).

The influence on evaluators to match their ratings to VAM outcomes was

observed in Tennessee where the TNDOE (2012) study noted an "inability or unwillingness" of evaluators to rate teachers at low levels (0.2% of teachers were rated at Level 1) compared to that of TVAAS® (where 16.5 % of teachers were rated at Level 1). Furthermore, Amrein-Beardsley and Collins (2012) showed the possibility of principal ratings being skewed to match multiple measures:

> In addition, some suggest that their supervisors are skewing their observational scores to match their SAS® EVAAS® scores given external pressures to do so (Collins, in progress). Such practices have been shown to occur elsewhere with the Tennessee Value-Added Assessment System (TVAAS) from which the SAS® EVAAS® was derived (Garland, 2012). In New York as well, if teachers have two years of low value-added scores, the teachers are to be rated ineffective overall and terminated, regardless of what other measures (e.g., supervisor evaluation scores) indicate or disclose (Ravitch, 2012). Because these other measures are often perceived as less objective, it seems that measuring teacher effectiveness using value-added output is beginning to trump other indicators capturing what it means to be an effective teacher. (p. 5)

Rubin et al. (2003) noted the effects of using VAM within teacher evaluation systems to judge teachers by comparing a similar situation in judging the quality of doctors and hospitals studied by Green and Wintfield (1995). This was an example of Campbell's (1976) law in action.

Green and Wintfield (1995) reported doctors "gaming the system" by increasing their initial diagnosis of patient sickness to a higher risk in the hope of inflating their outcomes. Rubin et al. (2003) advised to use VAM estimates of teacher effectiveness

among comparable units of measurement (compare inner-city school to another inner-city school) in order to maintain fairness within teacher evaluation systems.

Soar et al. (1983) proposed early in the history of using outcomes from standardized tests of students to evaluate teachers that if the "teachers' future" depended on these scores, "it seems inevitable that the simpler objective will be taught at the expense of more complex ones" (p. 243). Soar et al. identified that like the doctors in Green and Wintfield (1995), teachers would attempt to fulfill quantitative, accountability demands by teaching lower skill sets to their students so they would perform well on a standardized test that would in turn be used to evaluate teachers.

Although MET (2010, 2012) advocated for multiple indicators of teacher effectiveness within teacher evaluation systems in order to overcome unintended consequences, Headden (2011), the TNDOE (2012) study, and Amrein-Beardsley and Collins (2012) have shown evidence that such indicators may still become polluted. Future research will be needed to measure how multiple measures of teacher effectiveness within teacher evaluation systems will affect teaching practices and teacher leadership.

In summary, researchers have uncovered at least five unintended consequences associated with using VAM within teacher evaluation systems in order to improve teaching practices and teacher leadership. The five unintended consequences have been purportedly tied to interjecting a competitive element within newly revamped teacher evaluation systems.

Nichols et al. (2005) saw these unintended consequences as the result of using quantitative data to measure social outcomes. These unintended consequences have the potential to undermine improving teaching practices and teacher leadership; however,

future research will be necessary to measure their influences.

Of the subsections and encompassing categories derived from pilot data and reviewing the literature, the investigation into the unintended consequences of using VAM within teacher evaluation systems and the effects on teaching practices and teacher leadership has been most notably in need of future research (Amrein-Beardsley & Collins, 2012; Collins, 2014; Nichols et al., 2005). Although there has been missing evidence and data pertaining to this topic, this was expected as most newly revamped teacher evaluation systems are currently in the beginning stages. This does not detract the importance of studying such a topic within a case study approach; however, it has been noted by authors that consideration of this topic will require constant attention in the future to keep abreast of new data and research (Amrein-Beardsley & Collins, 2012; Collins, 2014; Nichols et al., 2005).

The Effects of Formative and Summative Evaluation

Formative evaluation. When principals have observed and evaluated teachers consistently within the classroom setting, principals have been able to collect ample evidence and make quality judgments about teacher performance rather than arbitrarily assigning ratings to teachers (TNTP, 2010). Replacing traditional lines of teacher evaluation characterized by “infrequent,” “unfocused,” “undifferentiated,” “unhelpful,” and “inconsequential” outcomes (TNTP, 2010, p. 1), recently revamped teacher evaluation systems have been designed to deliver consistent (Marzano et al., 2005), clear (Sartain et al., 2011), differentiated (MET, 2012), helpful (TNTP, 2010), and usable results that hold principals and teachers accountable for student learning (McCaffrey, 2013).

High-quality rubrics and improved training throughout the 2000s has given

principals unprecedented ability to validly and reliably judge teacher performance (Danielson, 2007). These rubrics were initially developed and refined using student achievement data stemming from the NCLB Act and have been compared to teacher performance that has produced the highest level of student learning (Kane et al., 2011; MET 2010, 2012).

Historically and more recently, researchers, policymakers, and commentators have agreed that an important function of teacher evaluation systems has been to improve teaching practices and teacher leadership *during* the process of teaching and learning (Darling-Hammond & Adamson, 2013; Shinkfield & Stufflebeam 1995). This approach has been known as formative evaluation and has been marked by targeted, "in-the-moment" (Donaldson & Peske, 2010, p.11) feedback to teachers alongside professional development addressing specific instructional needs (Darling-Hammond & Adamson, 2013). Newly developed rubrics have given this approach a basis from which to work (Danielson, 2007).

Examples of formative evaluation within teacher evaluation systems were uncovered throughout this literature review and the effects on teaching practices and teacher leadership were pronounced (Donaldson & Peske, 2010; Kane et al., 2011; MET, 2010, 2012). Historically, teacher evaluation systems were given lip service providing little on-the-spot intervention for identifying and improving struggling teachers. The RttT initiative has spurred improvements toward forming teacher evaluation systems to improve teaching practices and teacher leadership (Darling-Hammond & Adamson, 2013).

Kane et al. (2011) specifically showed how the results of multiple measures within teacher evaluation systems could provide teachers specific feedback to improve

instruction across different subjects. While finding that principal ratings predicted student achievement, Kane et al. (2011) was able to identify and inform teachers about which performance standards (i.e., classroom management skills, thought-provoking questions) led to higher student achievement.

Kane et al.'s (2011) outcomes had a positive effect on teaching practices and teacher leadership. Principals could concentrate on paying particular attention to specific standards that were shown to increase student learning during preconferences, classroom observations, and postconferences. Teachers could also translate Kane et al.'s (2011) feedback into their future lesson planning.

The same course of action held true in reports across the MET project. The goal was to identify teacher effectiveness measures, incorporate such measures as standards within teacher evaluation systems, rate teachers based on those standards, and check validity using multiple measures including VAM outcomes and student surveys (MET, 2010, 2012). The outcomes of the MET project centered on helping teachers improve teaching practices and teacher leadership using scientific and statistical precision to determine what teacher needed help and where they needed help along a spectrum of teacher effectiveness measures. Teacher evaluation systems have historically attempted to follow this approach but lacked the necessary data (Darling-Hammond & Adamson, 2013).

Like Kane et al. (2011), MET (2010, 2012) identified specific standards on which teachers could improve during instruction. Both Kane et al. (2011) and MET (2010, 2012) gave teachers the ability to change their approach specifically related to standards that related directly to improving student outcomes. Kane et al. (2011) found that when teachers used quality "questions/discussion" techniques, they produced higher student

outcomes, while MET (2012) found that teachers who did not were rated lowest on corresponding standards across the various rubrics (CLASS, FFT, PLATO, MQI, UTOP) and did not improve student outcomes.

MET (2012) found this to be of concern and that formative evaluation could be the solution. MET (2012) found that specific professional development could be necessary to address the following standards where teachers received the lowest principal ratings across rubrics: "Instructional dialogue" (CLASS), "Analysis and problem solving" (CLASS), "Using questioning and discussion techniques" (FFT), "Communicating with students" (FFT), "Classroom discourse" (PLATO), "Student participation in meaning making and reasoning" (MQI), and "Questioning strategies" (UTOP) (p. 4).

In contrast, MET (2012) found that teachers excelled and were rated highly on standards involving classroom management across the various rubrics. Generally, research across Kane et al. (2011), Sartain et al. (2011), and MET (2010, 2012) indicated formative evaluation could center on improving the ability of teachers to ask thought-provoking questions that would lead to an investigation of an event (history), process (mathematics), or theory (science).

This outcome-based approach could yield positive results in the future for teaching and learning as it has in Tennessee (TNDOE, 2012). Principals in Tennessee directly attributed improvements in teaching and learning to formative evaluation where conversations about the teaching and learning process occurred (TNDOE, 2012). MET (2012) called this approach a "qualitative coaching conversation" (p. 8) based on specific teaching standards, student outcomes, and an inquiry into how teachers could improve.

Although the literature has historically and consistently viewed the observation

and evaluation of teachers as time consuming, the ability of principals to offer specific feedback to improve teaching practices and teacher leadership hinged on their ability to be involved at the classroom level observing, collecting data, and analyzing data (Bryant, 2013; Natriello, 1983). One unique property of MET (2010, 2012) was the ability for multiple raters to randomly observe teachers using Teachscape technology that involved a panoramic camera able to capture the teaching and learning process in real time. Such an approach would give observers flexibility in providing formative feedback to teachers and allow reliability checks among observers (MET, 2012).

To have video available for observers would also increase the ability for teachers to quickly receive feedback in a timely manner, a key ingredient to formative evaluation. The time period between observed teacher performance and the time between teachers receiving feedback has shown to matter–the longer the period the less effect feedback had on improving teaching practices and teacher leadership (Donaldson & Peske, 2010; Frase, 1992). Moreover, teachers could observe themselves teaching increasing self-reflection and self-evaluation, which has been an important part of formative evaluation (Peterson, 2000).

Whether or not it would be feasible to reduce operating responsibilities of principals so they could dedicate the time necessary to observation and evaluation demands was beyond the scope of this study. However, if teacher evaluation systems will improve teaching practices and teacher leadership in the future, the literature has made a pronounced call to ensure principals have the time necessary to have meaningful conversations and use outcomes from observation and evaluation rubrics to improve teaching practices and teacher leadership (Bryant, 2013; Fink & Resnick, 2001; Natriello, 1983; TNDOE, 2012). Blase and Blase (2000) posited the principal as an important

component of improving teaching practices and teacher leadership by having ongoing conversations with teachers who inspire them to self-reflect and improve.

But as it was a concern with Natriello (1983), it has not appeared that the burden of administrators as classroom observers has lessened. Based on the work of the TNDOE (2012) study and Bryant (2013), administrators have recently inherited more responsibility in observing and evaluating teachers making it difficult to carry out a formative type of evaluation.

Because of time demands and the needs of teachers to receive ongoing, targeted feedback from observers well acquainted with curriculum, peer observation has taken on a more formal role (Jacques, 2013). The PAR program utilized this approach by recruiting highly effective teachers to serve as consulting teachers to new teachers or veteran teachers who have struggled (Goldstein, 2007). The PAR program is unique because it placed the consulting teachers front-and-center as agents of change within a system that has concentrated on formative evaluation (Goldstein, 2007).

After surveying consulting and participating teachers in PAR, Goldstein (2007) identified many facets in which consulting teachers provided formative evaluation to improve teaching practices and teacher leadership and how PAR differed from traditional teacher evaluation systems. Goldstein (2007) carried out a single-case study using observations, interviews, and surveys among consulting teachers, participating teachers, and the panel overseeing PAR cases.

Participating teachers in PAR showed trust in their consulting teacher because the ultimate goal was improvement (Goldstein, 2007). There was also ongoing feedback, for which teachers participating in PAR showed particular enthusiasm because the feedback was from an expert in their subject matter (Goldstein, 2007). Participating teachers being

reviewed in the PAR program identified that their consulting teachers were able to give them a level of targeted, subject-driven feedback not available to them in the past by administrators in traditional teacher evaluation systems (Goldstein, 2007).

The level of professional development was meaningful as shown by Goldstein (2007), who specifically stated,

> What can be drawn from this study, however, is that people involved in PAR believed that meaningful professional development was taking place as a form of evaluation that looked very different from teachers putting on a special show and principals flying through with a checklist. (p. 490)

Curtis (2011) reinforced Goldstein's (2007) trust factor in the relationship between peer observers and teachers in Washington D.C.'s teacher evaluation system. Master educators have been purposely separated in their duties from instructional coaches and administrators in the D.C. teacher evaluation system while all parties involved carefully maintain separate records helping to establish reliability (Curtis, 2011). The vast majority of master educators have been hired outside the school system, which has meant there were no personal ties allowing for unbiased observations and formative feedback (Curtis, 2011).

More importantly, master teachers concentrated on partnering with new and struggling teachers in a way that made teachers feel comfortable while being observed and participating in postconferences (Curtis, 2011). Curtis (2011) described master teachers as improving teaching practices and teacher leadership because they were trusted and were experts in the content of their teaching fields and because they received in-depth training.

Alongside using peer observations to improve teaching practices and teacher

leadership within teacher evaluation systems, substantial research supporting newly designed observation and evaluation rubrics (Danielson, 2007; Marzano et al., n.d.; Stronge & Associates, 2013) has shown improvements in the ability of principals to accurately judge teacher performance (Jacob & Lefgren, 2008). Rather than simply checking a box indicating whether or not a teacher meets a satisfactory level in order to justify tenure or a continuing contract, the newly designed rubrics of Danielson, Marzano, and Stronge have concentrated on identifying the quality of teaching practices and teacher leadership of teachers and improving performance.

Teachers have shown a willingness to internalize an observation rubric if they have perceived the rubric has been built on valid components and reliable ratings across observers (Headden, 2011; Taylor & Tyler, 2011). Specifically, the internalization of a rating rubric has meant teachers will have surveyed its components, sought to understand them, and will attempt to fulfill the standards in future classroom observations in order receive higher ratings (Headden, 2011). Based on their research, Taylor and Tyler (2011) remarked that teachers have been shown to "prioritize behaviors that increase" (p. 7) their observation scores on related classroom observation rubrics.

If the correlation of rating rubrics to student achievement has been accurate, if teachers view the rating rubrics as valid and reliable (and internalize), and if teachers perceive the principal as observing and gathering a reliable amount of data to make accurate judgments, these outcomes could drive improvement in student achievement. This could be an important feature within teacher evaluation systems going forward and offering formative evaluation to improve student achievement.

Alongside research-based classroom observation rubrics, the use of VAM has been another tool administrators now have within the framework of formative evaluation.

Principals have been able to compare their qualitative observations and judgments with VAM outcomes to increase the validity and reliability of the ratings they assigned to teachers (MET, 2012). Although VAM has been viewed as a more summative type of evaluation (Nussbaum, 2012), principals have had the ability to use VAM outcomes from preceding years to help guide discussions with teachers in following years (Sanders & Horn, 1998; Silva Mangiante, 2011).

Sanders and Horn (1998) outlined how using TVAAS® as a diagnostic could inform teachers of their effectiveness across low-, middle-, and high-achieving students. Having this historic diagnostic information on hand would benefit teachers to extrapolate what was needed to reach all groups among incoming students.

Whether VAM data has been used diagnostically by principals to help guide teacher evaluations or used by teachers planning for future students, it was assumed that growth patterns of past students would serve as beneficial inferences meant to improve teaching practices and teacher leadership. In this sense, both situations would be part of a formative evaluation framework in helping teacher improve teaching practices and teacher leadership for incoming students.

The EVAAS® system has provided schools and teachers projected growth scores that could be used as part of a formative evaluation plan for principals and teachers (SAS, 2007). At the teacher level, projected scores for students have given teachers a preview of historical data for their presently enrolled students (SAS, 2007). Projected scores have provided teachers a look at trends of learning outcomes their students have experienced in the past (SAS, 2007). If a particular student has experienced an historical downward trend, this would allow specific attention be given to that student. On the aggregate, teachers could receive feedback from principals pertaining to the level of differentiated

teaching practice offered to these students.

Taylor and Tyler (2011) carried out a study with significant meaning given the context of classroom observation rubrics, VAM outcomes, and formative evaluation within teacher evaluation systems. Taylor and Tyler (2011) studied the effects of observation and evaluation ratings on midcareer teachers within Cincinnati's teacher evaluation system. The authors specifically investigated whether teachers undergoing a full, formal evaluation cycle would improve during the evaluation cycle and whether that improvement would be sustained (Taylor & Tyler, 2011). Taylor and Tyler's (2011) sample was novel in that it involved a sample of teachers where 83.4% had between 5 and 18 years of teaching experience and totaled N=105 teachers with a total of N=14,208 students.

Taylor and Tyler (2011) found that there was an improvement on average for math teachers during and after intensive classroom observation and evaluation schedules from the teacher evaluation system in Cincinnati. However, Taylor and Tyler (2011) found no significant effects on reading teachers experiencing the same treatment.

Taylor and Tyler (2011) found that teachers experiencing the full cycle of classroom observations and evaluations produced an increase on average of 0.072 standard deviations in math. For the years after the full cycle of classroom observations and evaluation, teachers produced an increase on average of 0.111 standard deviations in math and "would continue at least over the first few years following a teachers TES evaluation" (Taylor & Tyler, 2011, p. 28).

Taylor and Tyler (2011) hypothesized that feedback provided teachers by "high-performing, experienced teachers who are external to the school" (p. 3) and by administrators spurred "employee investments in human capital production" (p. 29). This

formative evaluation provided teachers helpful feedback and "ultimately [teachers] adjusted their behavior long run" (Taylor & Tyler, 2011, p. 26).

Formative evaluation was a centerpiece in Sartain et al.'s (2011) study. After administrators successfully differentiated among teachers based on performance, which rarely occurred historically (Weisberg et al., 2009), administrators offered higher levels of feedback than before in Chicago's traditional teacher evaluation system. Sartain et al. was significant and already analyzed in preceding subsections. Sartain et al. was also important in measuring the effects of Danielson's (2007) *Framework* on formative evaluation in the CPS.

CPS expected principals and teachers to hold conferences before and after classroom observations and expected the conversation to be centered on instruction (Sartain et al., 2011). Specifically, CPS wanted teachers and principals to talk about

- How the lesson relates to the curriculum and the sequence of learning for the class.
- Characteristics of students in the class and how their individual needs varied.
- The goals for student learning.
- How the teacher will engage students.
- How the teacher will differentiate instruction.
- How the teacher will assess learning.
- If and how the teacher departed from the lesson plan.
- What changes the teacher would make if he/she could re-teach the lesson. (Sartain et al., 2011, p. 22)

Sartain et al. (2011) made the following remarks after surveying principals and

teachers and based on the authors' observations of conferences between principals and teachers in the CPS:

- Principals and teachers thought the conferences they had about instruction using Charlotte Danielson's Framework for Teaching were:
  - More reflective than those they had using the CPS checklist
  - Based on a shared language about instructional practices and improvement
  - Evidence-based, which reduced subjectivity
- Positive attitudes about conferences were dependent on principals' skills and buy-in.
- Our observations of the conferences revealed that the quality of the conversations could be improved and that principals need more support in engaging in deep coaching conversations. Conversations were:
  - Dominated by principal talk
  - Driven by low-level questions, although this varied across principals and teachers. (Sartain et al., 2011, p. 21)

Both Taylor and Tyler (2011) and Sartain et al. (2011) were significant because they showed quantitatively (Taylor & Tyler, 2011) and qualitatively (Sartain et al., 2011) that Danielson's (2007) *Framework* led to improvements in teaching practices and teacher leadership within a framework of formative evaluation. Both Taylor and Tyler and Sartain et al. also displayed the importance of teacher self-reflection, setting professional goals, and an overall theme of intrinsically motivating teachers–particularly Sartain et al.

Sartain et al. (2011) attested to principals' efforts in sustaining high-level conversations that caused teachers to develop a sense of personal responsibility in improving teaching practices and teacher leadership within a framework of formative evaluation. Sartain et al.'s findings stood in contradiction to Weisberg et al.'s (2009) where 75% of teachers communicated they had not received feedback on areas of improvement in their performance. This has been an ongoing aim of teacher evaluation for decades–to develop a sense of personal and collective responsibility to provide the best teaching practices and teacher leadership without the necessity of accountability measures (Darling-Hammond & Adamson, 2013; Milner, 1991).

Bandura (1977) posited self-efficacy as an individual's sense of personal responsibility to carry out a task and believing oneself capable to perform at mastery levels to complete the task, regardless of potential pitfalls. In the 1980s and 1990s, Bandura sought to discover what elements shaped a person's self-efficacy (Bandura, 1997). Many of Bandura's (1977) insights have been interwoven into the framework of formative evaluation to instill a teacher's self-efficacy.

For instance, Bandura (1977, 1997) found that incidences of success build self-efficacy, which means that administrators have had the ability to celebrate and highlight successes of teachers within the feedback loop described by Sartain et al. (2011), Taylor and Tyler (2011), and MET (2012). This has created a winning streak whereby a teacher's self-efficacy increased and teachers sought successes in other areas, perhaps some areas as guided by an administrator (Sartain et al., 2011).

Administrators also have had the ability to offer verbal cues (Bandura, 1977, 1997) to teachers in coaching situations that have invited teachers to self-reflect and think about various teaching practices and teacher leadership that suited a particular need.

Bandura's (1997) proposition appeared within a rubric designed by Sartain et al. (2011) to measure the level of questions proposed by a principal during conferences held after a classroom observation for teachers in the CPS. Sartain et al. used this rubric to categorize 300 principal pre and postobservation conferences involving 21 teachers and their principals into low, medium, and high levels of questioning techniques. On Sartain et al.'s rubric, the highest level of questioning used by principals reflected Bandura's (1977, 1997) proposition: "Principal's question requires extensive teacher response. The question and response reflect high expectations and require deep reflection about instructional practice. The principal and teacher push one another's interpretations" (Sartain et al., 2011, p. 24).

Bandura's (1977, 1997) insights were borrowed within the education realm and applied in formative evaluation where principals have been expected to model expected teacher behaviors on Sartain et al.'s (2011) rubric. MacNeill (2007) found that when principals offered modeling strategies within feedback, as long as the modeling tapped into teacher requisite skills and was not overcomplicated, it enhanced teacher self-efficacy. This sort of modeling has occurred in discussions during pre and postconferences using Danielson's (2007) *Framework* (Sartain et al., 2011).

Bandura's (1977, 1997) transition naturally flowed into a mental state of being–as employees (teachers) have become sure of their ability to make a contribution, they have performed at higher levels for personal satisfaction rather than meeting the needs of a rubric or accountability measures. Frase (1992) recognized principals as transformational leaders and commented with the following observation:

> Transformational leaders in schools provide a workscape where psychic rewards are more likely to be found, where people stand the best opportunity to motivate

> themselves successfully. They foster teachers' sense of ownership in the school mission and a strong belief in its importance. They help teachers find the crucial and natural link between job and personal satisfaction. (p. 180)

In summary, formative evaluation has had a powerful impact on improving teaching practices and teacher leadership within teacher evaluation systems (Kane et al., 2011; MET, 2010, 2012; Sartain et al., 2011; TNDOE, 2012). Newly designed teacher evaluation systems have been designed for the purpose of differentiating among teachers based on performance on multiple measures and offering support to low and middle performing teachers while replicating the teaching practices and teacher leadership of high performers (Goe, 2008).

Principals have played a central role guiding instruction at the school level (MacNeill, 2007). Kane et al. (2011), Sartain et al. (2011), the TNDOE (2012) study, and MET (2012) all showed that principals have used formative instruction to improve teaching practice and teacher leadership, both quantitatively and qualitatively.

One aspect requiring future research could involve an investigation into how newly designed teacher evaluation systems have influenced teachers to professionally grow based on their own initiative. The investigation could borrow methodology from Sartain et al. (2011) while attempting to uncover whether or not newly designed teacher evaluation systems have fostered Bandura's (1997) work.

Summmative evaluation. While a prominent use of newly designed teacher evaluation systems has been to differentiate among teachers based on their performance in order for formative evaluation to take root (TNTP, 2010), policymakers and researchers have also been interested in using teacher evaluation systems within the framework of summative evaluation (Peterson, 2000; Popham, 1988, 2013). Teacher

evaluation systems have been used historically within legislative mandates to employ a hybrid form of formative and summative evaluation (Popham, 2013).

While formative evaluation has centered on improving teaching practices and teacher evaluation, summative evaluation has concentrated on accountability measures (Peterson, 2000). Popham (1988) viewed summative evaluation as a tool to dismiss incompetent teachers if they could not improve. According to Popham (2013), Scriven (1967) began discussing the differences between formative and summative evaluation and viewed formative evaluation as therapy and summative evaluation as a diagnosis.

Whereas formative evaluation has been considered low-stakes mainly having been used to drive improvements in teaching practices and teacher leadership, summative evaluation has been based on high-stakes decisions about teachers, principals, and schools (Peterson, 2000). Using teacher evaluation systems to make high-stakes decisions has included dismissing low-performing teachers, granting or denying tenure to novice teachers, or determining which teachers to lay off when personnel has been reduced (Peterson, 2000).

Popham (1988, 2013) called the combination of formative and summative evaluation a "dysfunctional marriage" and has made the case for over 2 decades that combining processes within teacher evaluation systems for both formative and summative evaluation has failed. Popham (2013) contested that the underlying problem with combining processes for both formative and summative evaluation has been that administrators take on two roles–one as a mentor who has wanted to help teachers to improve and the other as a manager with the potential of dismissing teachers. Popham (2013) related the idea that teachers have wanted to improve teaching practices and teacher leadership, but teachers have been nervous about identifying improvements in

teaching practices and teacher leadership for fear of being dismissed.

Popham (1988, 2013) proposed the solution was to separate the administrative and observer roles in order to give teacher evaluation systems the ability to improve teaching practices and teacher leadership. McGreal (1983) noticed that summative evaluation had a "threatening nature" (p. 116) and that combining formative and summative evaluation "places both the supervisor and teacher in roles that inhibit their ability to operate in an open, cooperative manner" (p. 38).

Peterson (2000) blamed the bimodal function of administrators as facilitators of both formative and summative evaluations for "over 70 years of research studies that show the inaccuracy of administrator reports of teacher performance" (p. 6). Peterson specifically noted the negative effects on teaching practices and teacher leadership in relation to combining formative and summative evaluation into the role of an administrator:

> Administrators have a basic role conflict of interest when they are both summative judges and instructional leaders within the same population of teachers. Their perceptions and biases are shaped by their role assignment in the school. Sociologists describe a delicate balance of support and control that administrators require from educators in a school; the result is that evaluation activity and decisions become a tool for overall administrator functioning rather than an accurate, informative, and useful report and judgment of teaching quality. Administrator expertise in subject matter content is not as strong as other teachers in the same assignment. (p. 6)

A fundamental philosophy and design of the Peer Assessment and Review (PAR) has been to separate the roles of evaluation in teacher evaluation systems (Koppich,

2009). The PAR system has placed the formative evaluation responsibilities in the hands of consulting teachers who are peer experts in subject matter, independent of the school, and have concentrated on improving teaching practices and teacher leadership of teachers without a substantial aim of summative evaluation (Koppich, 2009).

However, the PAR system has also shown to be effective in summative evaluation. Koppich (2009) carried out a case study using the original PAR system in Toledo, Ohio. Koppich was significant because of its in-depth investigation into the PAR process, specifically the ability to summarily dismiss pretenured teachers who did not display quality teaching practices and teacher leadership. Koppich also looked at the PAR system's ability to dismiss incompetent veteran teachers when they failed to improve their teaching practices and teacher leadership.

While many authors have agreed that it has been difficult to dismiss incompetent teachers, especially veteran, tenured teachers (Peterson, 2000; Weisberg et al., 2009; Wise et al., 1984). Koppich (2009) found that on average the PAR system in Toledo successfully nonrenewed 8-10% of first year teachers between the years 1981-1982 and 2007-2008. Koppich noted this total did not include participating first-year teachers who resigned before the end of the year knowing they would not proceed through the PAR process. Also, Koppich showed that of the 18 veteran teachers placed in the PAR system between the 2004-2005 and 2007-2008 school years, five were dismissed or resigned, 10 continued on in the program, and three were successfully returned to the classroom.

This outcome stands in contrast with Weisberg et al.'s (2009) research where of five public school systems, an average of 0.1% to 0.9% of beginning probationary teachers were nonrenewed for performance reasons each year. Denver Public Schools was an outlier in Weisberg et al.'s (2009) study with an average nonrenewal rate of 3% of

beginning probationary teachers for performance reasons, which was still less than Koppich's findings.

Goldstein (2007) observed the summative nature of the PAR system in "stark contrast to the automatic granting of tenure that often meets new teachers after a set number of years of service" (p. 496). Goldstein and Noguera (2006) elaborated with the following:

> When a district undertakes the systemic changes involved with peer assistance and review, the quality of teacher performance becomes a central concern. Losing 10 percent of a cohort of new teachers as a result of rigorous evaluation is virtually unheard of in urban districts across the United States. It is even rarer to find a district that removes underperforming veteran teachers without a high degree of conflict and rancor. Districts that have successfully implemented peer assistance and review have found it an effective means to systematically improve the quality of teaching and, in the process, to honor and recognize the best teachers. (p. 36).

Curtis (2011) documented the effectiveness of summative evaluation used within Washington D.C.'s teacher evaluation system already analyzed in preceding sections. IMPACT has made use of master educators who have been parallel to consulting teachers in PAR systems.

Curtis (2011) identified master educators as "important partners" (p. 14) in IMPACT and showed summative data that supported the ability of IMPACT to differentiate among teachers based on performance. Because of principal and master educator classroom observations alongside VAM measures, 3% of teachers were found to be "ineffective" and 16% "minimally effective." This outcome was vastly different than

the year preceding IMPACT where 0.2% where identified as “unsatisfactory” and 4.8% as “needs improvement” (Curtis, 2011).

Curtis (2011) noted that if teachers rated “ineffective” did not improve their teaching practices and teacher leadership, they could be dismissed within a year; while teachers rated “minimally effective” had 2 years to improve their performance. In 2010, the D.C. public school system dismissed 125 teachers who were rated “ineffective” in the fall of 2009. At the time of the Curtis study, teachers rated “minimally effective” were working to improve their performance.

While VAM was used in D.C. within IMPACT to help determine dismissals, VAM has seen an uptick in its use within teacher evaluation systems for summative evaluation (MET, 2012; Sartain et al., 2011). As previously noted, VAM has been primarily viewed as an approach within a framework of summative evaluation (Nussbaum, 2012).

In its present use, VAM has been labeled a tool of summative evaluation with some formative features (Nussbaum, 2012). Within the mode of summative evaluation, many authors recommend that VAM should not serve as a singular source of data for judgment of teaching practices and teacher leadership within teacher evaluation systems (Glazerman et al., 2010; Haertel, 2013; MET, 2010).

North Carolina has adopted this advice to an extent in that there are two primary sources of data used to measure teaching practices and teacher leadership–principal ratings and EVAAS® outcomes (McREL et al., 2012). However, within the design of NCEES, it will be possible in the future for a teacher to be rated “developing” on Standards 1-5, or a teacher to be rated “does not meet expected growth” in Standard 6, and be found “in need of improvement” for an overall status (McREL et al., 2012).

There are two important features within NCEES–yearly ratings and a yearly status. The yearly rating has been designed to be a within-year judgment of teaching practices and teacher leadership that could be used to help teachers grow through formative processes such as post-conferences with principals and professional development (McREL et al., 2012). The overall status of a teacher refers to a final, summative judgment about a teacher's standing and, for EVAAS® outcomes, will be determined after a 3-year rolling average has been calculated. A recent change will consider an average of the two highest EVAAS® results over the course of 3 years while dropping the lowest year.

The important point within this context has been that NCEES does incorporate multiple measures of a teacher's performance using principal judgments in Standards 1-5 and EVAAS® outcomes in Standard 6, but NCEES seemed to not incorporate multiple measures equally. It seems possible in the future that any one summative indicator could trump the others, which seems to contradict advice from the literature. For example, a teacher could receive "proficient," "accomplished," or "distinguished," or a combination thereof on principal ratings from Standards 1-5 yet not "meet growth" expectations in Standard 6. This teacher could be found "in need of improvement" for a summative yearly status if EVAAS® results were not meeting expectations for at least 2 or more years in a row, regardless of their ratings on Standards 1-5 each year.

Although it seems as though NCDPI has been conservative in its summative approach in requiring 3 years of EVAAS® data and taking an average of the two highest years of EVAAS® results, the possibility exists that one standard could trump another and conflict with using multiple measures of teacher effectiveness within NCEES, at least in

the sense of using them equally. It is important to note that at the time of this study, teachers have not received 3 years' worth of data within NCEES, so no teacher in North Carolina has received a final summative status.

Some states have used a point system where ratings for each standard are worth points towards a total that the teacher accrues towards a final, summative evaluation (Miami-Dade County Public Schools [MDCPS], 2015). This approach has disallowed any one measure to trump another because they have counted equally towards a summative total. As long as teachers reached a cut point, teachers have been considered proficient or higher in a summative sense.

Nussbaum (2012) provided an intricate look into using VAM in formative evaluation versus within summative evaluation. Nussbaum's analysis was novel because it gave a present-day account of using VAM in a legal sense and provided possible legal struggles ahead, given the newness of VAM being used formally within teacher evaluation systems.

Nussbaum (2012) questioned VAM used within summative evaluation based on its complexity and unreliable results. The following statement summed up Nussbaum's position:

> The problem of model prediction error can be mitigated when the value-added statistic is used in a formative evaluation. The statistic itself will mislabel 25% of "average" teachers as below average.[14] It is not disputed that there are a number of variables outside the control of the teacher that affect the value-added number. There will be too many overreactions on the part of the school district, about a teacher's abilities if the value-added number is used in a summative evaluation. In the formative evaluation, school districts will be able to examine the figures and

> determine whether a particular teacher's value-added number is an anomaly or a trend. This examination will allow for honest dialogue between a teacher and the school district because they do not have the imminent concern of being fired. The hardened stance of a teacher being fired after 2 years of a dissatisfactory value-added number is far too harsh. Accountability is crucial in creating good teachers but, rushing to judgment about a teacher because of a statistic that is not foolproof is just bad policy. (pp. 10-11)

In summary, formative and summative evaluation have been delineated. The literature has been robust in addressing how formative evaluation has positively affected teaching practices and teacher leadership within teacher evaluation, both historically and in the present day (McGreal, 1983; Popham, 1988; Sartain et al., 2011; Taylor & Tyler, 2011). However, there has been little evidence of summative evaluation improving teaching practices and teacher leadership within teacher evaluation and has served as an accountability piece for the general public in judging the overall quality of teachers (McGreal, 1983).

Taylor and Tyler (2012) showed the ability of formative evaluation quantitatively to improve teaching practices and teacher leadership. Sartain et al. (2011) supplied evidence that formative evaluation can be a powerful qualitative tool for principals in building the self-efficacy of teachers. In many cases, formative evaluation has shown to improve teaching practices and teacher leadership through in-the-moment feedback from principals and peers (Donaldson & Peske, 2010; Headden, 2011; Sartain et al., 2011).

McGreal (1983), Peterson (2000), and Popham (2013) outlined the tension between using teacher evaluation systems for improvement and accountability. This debate started many decades ago and has been unresolved.

Chapter 3: Methodology

Introduction

The topic under investigation in this study explored the effects the recently implemented performance-based teacher evaluation system in North Carolina has had on teaching practices and teacher leadership in the support of student learning. The topic has a high rate of relevancy due to the federal RttT initiative and because new teacher evaluation systems have been introduced at the state level across the country. It is imperative to analyze the effects of teacher evaluation systems in order to understand the impact on teachers and students (Amrein-Beardsley & Collins, 2012; Collins, 2014).

The basis for the choice of setting in this study was rooted in the modern high school reform movement largely affected by the literature in *Breaking Ranks: Changing an American Institution* (National Association of Secondary Principals [NASSP], 1996). This work was authored by NASSP with help from the Carnegie Foundation for the Advancement of Teaching.

An important factor in choosing teacher evaluation systems in high schools rested on a logical connection that by better observing and evaluating teachers and helping them improve their teaching practices and leadership roles, high school reform could be better established. High school reform efforts that began in the 1990s have continued to escalate as high school officials scramble to investigate successful approaches to turning high schools around as McNeil (2003) noted,

> For over two decades, the modern school reform movement has included efforts by commissions, reformers and researchers to address the ills of the American high school. The litany of shortcomings is long and well documented. On almost every statistical measure and for large groups of students, our high schools are not

> making the grade. At a time when the needs of our youth and the demands of society, the workplace, and life have changed dramatically, high schools have not been able to respond. Graduation rates have hovered around 75 percent for about 30 years. The performance of 17 year olds in reading is down. Achievement of 17 year olds in math and science is up, but those gains are largely attributable to improvements between grades 5 and 8, not from gains during the high school years. And, math and science achievement falls below that of young people in most developed Countries. While the gap in achievement, graduation, and college attendance between white high school students and minorities narrowed during the 1970s and 1980s, it widened again in the 1990s and the trend is continuing. (p. 11)

McNeil also added,

> For all concerned, it has been easier to focus attention on early childhood and K-8 education and hope that if we get those years right, the problems in our high schools will take care of themselves. But even though elementary and middle schools are sending students to high school with higher levels of reading, math and science skills, performance at the high school levels is not improving. One explanation: Like younger children, today's high school students have a different set of needs and different kinds of experiences than they did 50 years ago, when the current high school design was formulated. (p. 11)

Lachat (2001) pointed out that one important factor for overcoming barriers to high school reform was breaking the strong allegiance to the status quo in teacher evaluations. The status quo for teacher evaluation systems continued into the 2000s until the RttT initiative was introduced (TNTP, 2010).

Also, Markley (2006) noted several criticisms of evaluations resulting from a study by NCES (1999), one of which was the lack of expertise in teacher evaluations at the secondary level. One path to meaningful high school reform has been as Lachat (2001) directed: "A new mission for education that requires high schools to not merely deliver instruction, but to be accountable for ensuring that educational opportunities result in all student learning at high levels" (p. 10). One component of holding high schools accountable within the high school reform movement would be observing and evaluating high school teachers using a teacher evaluation system that promotes effective teaching practices and teacher leadership roles.

Therefore, in response to the high school reform movement in NASSP (1996), McNeil's (2003, p. 11) observation of stagnant teacher evaluation methods in high schools, Markley's (2006) observation of a lack of expertise in evaluating teachers at the secondary level, and the findings from TNTP (2010) that teacher evaluation systems have been in need of innovation, the focus of this research project was in six high schools in the southwestern corner of the piedmont region in North Carolina. The accessible population from which the sample was drawn included all of the teachers and administrators within these six high schools. The target population for which the results could be generalized included counties in North Carolina or other states containing similar composition and staffing as this piedmont county and that have undergone similar observation and evaluation processes.

Multiple sites including six high schools within this piedmont county were involved within the research design in this single case study, making the investigation similar to a multiple case study. However, it was defined as a single case study with "embedded units," as the "case" involved relates to NCEES and the "units" involved are

the six high schools (Baxter & Jack, 2008, p. 550). The advantage of the single case study with embedded units was that because the case (NCEES) was studied across multiple sites (six local high schools), the outcomes should be more robust in keeping with a multiple case study (Baxter & Jack, 2008).

All six high schools received equal treatment in completing a survey process and in carrying out interviews with principals. The data were analyzed in a parallel manner across the cases (school sites) in order to uncover similar ideas and concepts that may potentially lead to theories in determining the effects of NCEES on teaching practices and teacher leadership roles in support of student learning.

The framework for this research project lies is in a single case study format with embedded units, and grounded theory was used concurrently throughout the research project. Grounded theory has already been used to code qualitative data collected through pilot data from two focus groups and two principal interviews. This was in keeping with Charmaz's (2006) approach in delaying the literature review until collecting baseline data to help guide research. The results of the grounded theory analysis generated categories which were refined with a review of the literature both historically and in the present day.

The single case study research method with embedded units helped enforce the boundary of the research topic from a top-down perspective, while grounded theory inductively produced core concepts through the coding of qualitative data from a bottom-down perspective. The grounded theory aspects of the research should verify extant theories produced from the case study framework including congruence with data presented in the literature review.

A case study has been defined by Yin (2009) as "an empirical inquiry that investigates a contemporary phenomenon in depth and within its real-life context,

especially when the boundaries between phenomenon and context are not clearly evident" (p. 18). Charmaz (2006) defined grounded theory as "flexible guidelines for collecting and analyzing qualitative data to construct theories 'grounded' in the data themselves" (p. 2).

Combining these two research approaches helped validate findings in that if similar theories were produced from the top-down processes of the single case study approach compared to the bottom-up processes of the grounded theory approach, the conclusions were further validated. Fernandez (2005) accounted for Glaser's (1998) attestation that case study and grounded theory approaches could be hybridized (Chapter 5, online document). Fernandez provided many functional examples where researchers combined case study and grounded theory approaches to produce positive outcomes. According to Fernandez, Lehmann (2001, p. 87) claimed that

> Applying Grounded Theory to Case Study was very successful. It produced a prolific amount and yielded a great richness of information. . . . The case settings, furthermore, contained more varied data than could be expected from individual, purely homocentric studies. Efficiency and abundance combined to make this method an exceedingly fruitful one. (Chapter 5, online document)

Flick (2009) specifically dealt with combining research methodologies in order to avoid the "restricting subscription to a specific methodological discourse" and labeled the combination of research methods as "hybridization" (p. 459).

This single case study with embedded units incorporated a mixed-methods (QUAL-Quan) approach for data collection and treatment of variables in an attempt to study the effects of NCEES on teaching practices and teacher leadership in support of student learning. According to Gerring (2007), case studies "employ a great variety of

techniques–both quantitative and qualitative–for the gathering and analysis of evidence" (p. 33).

This single case study with embedded units provided a "thick description" (p. 49) of real-world events involving the teacher evaluation process and how it supports student learning in terms of teaching practices and teacher leadership roles (Gerring, 2007). Grounded theory helped build theories from qualitative data collected in the form of unstructured items on a questionnaire for teachers and from interviews with administrators who were involved in NCEES. This application of grounded theory corresponded with Charmaz's (2006) in that the "data form the foundation of our theory and our analysis of these data generates the concepts we construct" (p. 2).

Data Collection

Number of participants. Because the goal was to measure the effects of NCEES on teaching practices and teacher leadership of the teachers in high schools in the county under study, the entire population of high school classroom teachers in the six high schools in this county underwent survey procedures. The population was entirely accessible and at the time estimated to be small enough (estimated N=300) to access email addresses for survey procedures. Only classroom teachers were part of the population; this excluded counselors, therapists, nurses, and teacher assistants. Any classroom teacher who received observations and evaluation under NCEES during the 2014-2015 school year was considered part of the population under study.

Unstructured items were also part of the questionnaire in order to receive qualitative feedback to be analyzed using grounded theory. In favor of receiving as much qualitative data as possible, it was advantageous to give all teachers the ability to provide input. Gay et al. (2009) advised that for smaller populations (N=100 or fewer), the entire

population should be surveyed; but for populations larger than N=500, 50% could be sampled in various ways including randomly, stratified, clustered, or systematically. Since the population of teachers in the six high schools in the county under study was smaller than N=500 and email addresses of potential participants were accessible, the entire population was surveyed.

For principal interviews, participants were chosen more purposefully. The criterion for the "qualitative sampling" (Gay et al., 2009, p. 135) of principals or assistant principals rested on the principal's prior exposure to teacher evaluation systems. The goal was to purposely choose principals who had observed and evaluated teachers using other teacher evaluation systems than only NCEES. This step was to ensure that principals could accurately identify any possible effects of NCEES that were novel in nature compared to the historical systems within which principals had operated.

Demographics of sample. The most prominent demographic to be profiled and tabulated in analyzing survey data was teachers who have experienced receiving classroom observations and evaluation under teacher evaluation systems other than NCEES. A specific question was asked of survey participants regarding this status at the beginning of the survey and was tracked (see Appendix A, Survey Question 1).

This subgroup was important in determining the effects of NCEES versus effects of prior teacher evaluation systems that teachers had experienced either in North Carolina, in other states, or even different countries. The perceptions of teachers who have received classroom observations and evaluation using other teacher evaluation systems were statistically tested against teachers who had only experienced NCEES. Teachers who had experienced classroom observations and evaluation under other teacher evaluation systems were labeled "Teachers+1" for the sake of analysis; and

teachers who had only experienced NCEES were received the label “Teachers+0.” Teacher responses to this question were tracked for use in later statistical tests.

The surveyed population was also profiled based on years of teaching experience, ethnicity, and gender. A specific question was asked of survey participants regarding the status of each of these demographic profiles at the beginning of the survey (see Appendix A, Survey Questions 2-4). The responses of survey participants to these questions were tracked in order to carry out statistical tests.

This profiling (for Survey Questions 2-4) was a logical step extending from the work of Lynn et al. (2013). After running regression estimates for large teacher samples (N=10,616) (see Appendix U1 for reference), Lynn et al. found that teachers with 11-25 years of experience, men, and Black teachers were more likely to receive higher principal ratings for which EVAAS® scores were lower when comparing groups across similar NCEES ratings. Lynn et al. made the following statements in their findings:

> Male teachers with ratings for Standards 1 through 5 that are similar to those for female teachers also tend to receive lower teacher value-added scores. Similarly, Black teachers with similar ratings from principals receive lower teacher value-added scores than do White teachers. These findings suggest that male teachers, Black teachers, and teachers with 11 to 25 years of experience receive relatively higher principals’ ratings than value-added scores. (pp. 24-25)

The specific number of survey participants with their years of experience was listed alongside other demographics (gender and ethnicity) described in data tables using descriptive statistics. This allowed the opportunity to test the perceptions of teachers within each demographic subgroup for outcomes such as those found within Lynn et al. (2013). All demographic responses were tracked and displayed in tables for use in

various statistical tests.

Forms of Data Collection and Uses

Data collection involved three sources: survey data, principal interviews, and data from NCDPI for quantitative analysis. First, the survey process provided the ability to analyze teacher perceptions singularly for each item (including mean, median, mode, and standard deviation) on the questionnaire in Appendix A. The questionnaire also included an analysis combining ratings at both extremes, "strongly disagree" and "strongly agree," into two general categories of "disagree" and "agree." The outcomes of both the singular and combination analyses were provided in data tables and each included a summary of central tendencies for each item on the questionnaire.

The questionnaire also generated data from Questions 7-41 (see Appendix A) that were tested against demographic data from Questions 1-6 (see Appendix A) to form the basis of an inferential statistical analysis. This treatment allowed an analysis of how teachers viewed NCEES who had experienced classroom observations and evaluation under other teacher evaluation systems. It also investigated the perceptions of teachers who had only experienced classroom observations and evaluation using NCEES. The outcomes of the inferential statistics were also provided in a data table.

In projecting small responses across the five-point Likert scale, the Mann-Whitney U test was used to determine if there was a statistically significant difference between those teachers who had received classroom observations and evaluation under at least one other teacher evaluation system (Teacher+1) compared to those teachers who had only received classroom observations and evaluation using NCEES (Teacher+0). The following null and stated hypotheses were given and applied to Questions 5-22 individually on the NCEES Teacher Survey (Appendix A):

H_0=There is no statistically significant difference between Teacher+1 and Teacher+0 in their calculated mean perceptions on the NCEES Teacher Survey;

H_1=There is a statistically significant difference between Teacher+1 and Teacher+0 in their calculated mean perceptions on the NCEES Teacher Survey. (α=0.05)

As Cohen et al. (2007) advised, the nonparametric Mann-Whitney U test would not identify where the differences were between the groups Teacher+1 and Teacher+0, if there were any. If it was necessary to investigate differences based on significant p values, the researcher referred to tables and graphs to identify trends in an attempt to locate at which end of the Likert scale the differences were located for each group.

The same statistical approach was taken to test the perceptions based on years of experience, gender, and ethnicity. However, testing the perceptions of Teachers+1 versus Teachers+0 deserved a special status in this study because this step indirectly explored the effects of NCEES on teaching practices and teacher leadership while considering teachers who had experienced at least one other teacher evaluation system. When the group Teacher+1 responded to survey questions, their perceptions were at least in part built on what they have experienced previously within traditional teacher evaluation systems.

Data tables were built using all four demographic groups (Teacher+1/Teacher+0, teacher experience, male, and Black). For teacher experience, male, and Black teachers, the following null and stated hypotheses were combined for the sake of ease in the following and applied to Questions 5-22 individually on the NCEES Teacher Survey (Appendix A):

H_0=There is no statistically significant difference between teacher experience,

male, or black teachers in their calculated mean perceptions on the NCEES Teacher Survey;

H_1=There is a statistically significant difference between teacher experience, male, or black teachers in their calculated mean perceptions on the NCEES Teacher Survey.

(α=0.05)

The questionnaire used within the survey process also included unstructured items. This approach gave teachers the opportunity to provide qualitative data analyzed using grounded theory. This also allowed the opportunity to tease out data concerning the two aforementioned groups, Teacher+1 and Teacher+0. If there was a significant inferential difference, using grounded theory to explore responses to unstructured survey questions may have uncovered possible differences.

A second source of data was collected from principals who have operated within at least one teacher evaluation system other than NCEES. The principals were interviewed using the Principal Interview Questions in Appendix T1. The questions were created while reviewing the literature in the secondary research and from refining questions used for two principal interviews from pilot data. Grounded theory was used to code and analyze principal responses to the interview questions.

A third source of data was NCDPI's (2014) NCEES database found online. Within this third source of data, many statistical tests were used. First, the North Carolina Teacher Summary Ratings (NCTSR) was analyzed for the county under study. Bar graphs were constructed using NCTSR for the county under study. These bar graphs showed the composite frequency of teachers rated at each level using the NCDPI NCEES database, and the results were compared side-by-side to bar graphs that were constructed

from the following three observed sources: Weisberg et al.'s (2009) "Two Districts With a Four Point Rating Scale" (see Appendix B), Curtis's (2011) "SY 2009-10 Teacher Impact Ratings" (see Appendix D), and Lynn et al.'s (2013) "Number of Teachers in Each Rating Category by Standard" (see Appendix F). A descriptive analysis of the bar graphs followed.

Second, an inferential approach was taken using the aforementioned three observed sources. A chi-square goodness of fit analysis identified whether NCTSR from the most recent year of results for the county under study were significantly different from each of the three observed sources. The test statistic χ^2 was used to find "p" (probability) and describe how likely NCTSR in the county under study was different from each of the three observed sources.

The following null and stated hypotheses were given for each chi-square test involving the three observed sources:

H_0=There is no statistically significant difference between the NCTSR for the six high schools in the county under study and the three observed sources;

H_1=There is a statistically significant difference between the NCTSR for the six high schools in the county under study and the three observed sources.

(α=0.05)

The reasoning for testing NCTSR for the county under study against three observed sources was that the literature specifically found that newly revamped teacher evaluation systems specialized in distributing teachers across a broad range of performance standards (Batten et al., 2012; Daley & Kim, 2010; Lynn et al., 2013; Sartain et al., 2011). The RttT initiative was designed upon such an approach and included the mandate that stated, "judge (teacher) effectiveness using multiple rating

categories" (Lohman, 2010, para. 23). The approach to distribute teachers based on performance was in response to the historical skewed results of inflated principal ratings (Weisberg et al., 2009).

The last quantitative approach was taken in analyzing the year-to-year movement of NCTSR in the county under study. The process of analyzing teacher evaluation rating movements from NCTSR began from the 2011-2012 school year and continued through the 2013-2014 school year. The ratings from NCDPI in its annual NCTSR were used to formulate side-by-side bar charts based on contingency tables from the teacher evaluation ratings.

Contingency tables were built and based on the teacher ratings for each NCEES composite standard garnered from and dependent upon the yearly application of NCEES. The goal in comparing NCTSR was to identify improvements or declines in teacher ratings from year to year, ultimately providing a way to measure the success or failure of NCEES to improve teaching practices and teacher leadership.

For example, hypothetically if 5% of teachers were rated at "developing," 20% "accomplished," 50% "proficient," and 25% "distinguished" in 2011 for a specific standard and that rating changed to 2% "developing," 15% "accomplished," 48% "proficient," and 35% "distinguished" in 2012, certain conclusions could be drawn as to the impact NCEES instituted on teaching practices and teacher leadership roles. As stated in the NCEES handbook, "The instruments are designed to promote effective leadership, quality teaching, and student learning while enhancing professional practice and leading to improved instruction" (McREL et al., 2012, p. 4). It was a logical extension that if NCEES was designed to "promote effective leadership" and "quality teaching," NCTSR should improve year to year.

A chi-square goodness of fit analysis identified whether NCTSR from an expected year (for example 2011-2012) were significantly different from an observed year (for example 2012-2013). The test statistic χ^2 was used to find "p" (probability) and described how likely NCTSR had changed in terms of distribution based on chance or the implementation of NCEES. Each standard and objective outlined in the rubric by the NCEES handbook was subjected to the chi-square goodness of fit analysis. The following null and stated hypotheses were given:

H_0=No ascertainable difference in NCTSR across standards and years due to implementation of NCEES;

H_1=Ascertainable difference in NCTSR across standards and years due to implementation of NCEES.

(α=0.05)

In summary, NCTSR provided data used for quantitative analysis to show the distribution of teachers among composite ratings across each standard and whether that distribution was significantly different across three observed sources. Also, NCTSR was used to determine if there was significant movement of teachers across standards of the rubric in NCEES.

The need was also satisfied for the primary research to include an affective side of human behavior in terms of the attitudes and perceptions of teachers and administrators. Gable and Wolf (1993) defined affective characteristics as "attitudes, values, self-esteem, and interests" (p. 2). These characteristics were measured in teachers and administrators by administering the NCEES teacher survey with unstructured items and principal interviews. These qualitative data were analyzed using grounded theory coding processes and shed light on the attitudes of administrators and teachers about the impact of NCEES

on teaching practices and teacher leadership.

The multiple sources of data from the research allowed for triangulation within this case study which further substantiated conclusions drawn from analyzing data collected from the quantitative tests and the literature review (Bryman, 2001). The data triangulation (Gay et al., 2009) in this research project drew from quantitative data in analyzing NCTSR coupled with qualitative data stemming from unstructured survey questions for teachers and administrator interviews. Gay et al. (2009) defined triangulation as a

> Process of using multiple methods, data collection strategies, and data sources to obtain a more complete picture of what is being studied and to cross-check information. The strength of qualitative research lies in collecting information in many ways, rather than relying solely on one, and often two or more methods can be used in such a way that the strength of one compensates for the weakness of another. (p. 377)

Denzin (1989) categorized triangulation into four areas of which this study made use of varying data sources in terms of both the individuals supplying the data and how the data were acquired.

Assumptions of data collection and uses. There were many assumptions woven into the data collection and their uses in this study. The first set of assumptions applied to the survey research methods.

In the process of surveying a population, Gay et al. (2009) identified that it was possible that a certain demographic respond. It was possible that teachers at opposite ends of the spectrum responded: those who received negative feedback regarding their evaluation outcomes and were possibly disgruntled and those who received positive

feedback regarding their evaluation outcomes and were happy to share positive results. Also, according to Gay et al., response rates may be low. Both situations dampen generalizability to a target population.

Also, an assumption was made about the demographic group "Teachers+1" within the survey procedures. These teachers experienced receiving classroom observations and evaluation in at least one teacher evaluation system other than NCEES. It was assumed that these teachers represented a pre-NCEES and post-NCEES treatment. There was no way to validate this assumption entirely; rather it is assumed.

The collection and analysis of qualitative data in this study also created assumptions, specifically qualitative data from unstructured items on the NCEES Teacher Survey and from the NCEES Principal Interview questions. Some researchers have criticized the generalizability of qualitative data. However, Gay et al. (2009) identified qualitative data as important regardless of its level of generalizability. Gay et al. detailed qualitative data as descriptive and unable to provide ultimate truth, which has been an acceptable parameter of any research. Also, the analysis of qualitative data contains contextual power to the researcher and audiences experiencing similar phenomena regardless of the ability of qualitative data to uncover truth (Gay et al., 2009).

There were two important assumptions made within the quantitative analysis of this study that limits conclusions. First, in using NCTSR to describe the distribution of teachers across performance standards and using the three observed sources as a basis of comparison, assumptions were made about the similarities of the teachers within each group. However, the intent was to paint a broad brush stroke within the case study boundaries that illuminated the effects of NCEES on teaching practices and teacher leadership.

Second, although the four chosen sources of this analysis were used because of their relative importance in the literature, their rating schemes are not the same: They use different terminology and titles for each rating compared to each other and compared to NCEES. This design provided a simplistic picture of where the teachers in the county under study were compared to other groups in order to determine whether NCEES has successfully described teaching practices and teacher leadership across a broad range of teacher performance.

Second, in using the NCTSR to determine the movement of teachers across the rubric rating scale and across standards, an assumption was transferred about the stability of teacher employment across years. Teacher turnover has been high in education (Smith & Ingersoll, 2004). There was minimal chance that the population of teachers one year would be exactly the same the following year of the analysis. While this limited applicability, this analysis served to determine the effects of NCEES on teaching practices and teacher leadership on a broad, holistic scale.

Steps of Data Collection

1. Pilot data were collected using two focus groups and two principal interviews under the authority of the dissertation committee chairperson overseeing this study. The pilot focus group and principal interview questions were checked for face validity and were designed to be open-ended to allow free responses. There were no other validity concerns because of the desire to allow teachers and principals to talk freely about whatever they deemed important about NCEES and teacher evaluation in general. The researcher was interested in using this pilot data to solidify a choice for the topic of study, which at that point was somewhat undecided.

In the pilot data, the researcher allowed the discussion to go wherever the

participants guided it as long as the discussion involved processes surrounding teacher evaluation systems. This was not a formal investigatory endeavor.

The pilot focus groups and principal interviews were a way to collect initial, qualitative data to help guide the literature review. The pilot focus groups and principal interviews lasted for about 30 minutes each, and the recorded focus groups and principal interviews were personally transcribed and coded by the researcher. Charmaz (2006) advised delaying the literature review while using initial data for direction, which was the approach used here.

2. The literature review was then used to examine and determine the relative importance of the initial categories derived from the pilot data. The initial categories were successfully refined and displayed in Figure 2 of Chapter 2.

The initial categories from the pilot data included the quantity of classroom observations and evaluation, the quality of classroom observations and evaluation, the use of VAM to estimate teacher effectiveness, and how outcomes were used. These initial categories were generated from a grounded theory analysis of the pilot data and found to be important and relevant within the literature pertaining to the effects of teacher evaluation systems.

The literature review was replete with data that refined the initial categories. The literature review served to investigate the categories adding relevant data and conflicts in the literature, all of which were used to design a questionnaire for teachers and interview questions for principals in the primary methods of this research project. Each item on the questionnaire and each question used to interview principals can be traced directly back to the literature review. The content of the teacher questionnaire and principal interview questions was designed around the content and layout of the literature review which was

derived from initial categories from the pilot data. The items on the questionnaire and interview were placed in the same order as the content of the literature review.

Gay et al. (2009) referred to this approach as two-fold–to appropriate construct validity and content validity. Both of these types of validity were lacking in the instrumentation used to collect pilot data, specifically the focus group and principal interview questions–they were checked solely for face validity. However, both construct and content validity checks were apparent in the NCEES Teacher Survey and NCEES Principal Interview questions as evidenced in the literature review.

Gay et al. (2009) defined construct validity as the ability of an instrument to measure a hypothetical construct (p. 157) and content validity as the degree to which an instrument "measures an intended content area" (p. 155). The case study approach of this study provided fundamental boundaries in which the construct and content of the NCEES Teacher Survey and NCEES Principal Interview questions were derived. Specifically, the case study approach was bound by the research questions of the literature review including an exploration of the present-day and historical limitations of teacher evaluation systems and the effects of those limitations on teaching practices and teacher leadership (see Research Question 1 of literature review). Also, within the case study boundaries, present-day and historical processes of teacher evaluation systems were explored, alongside the effects of those processes on teaching practices and teacher leadership (see Research Question 2 of literature review).

Although not using formally validated instruments, the pilot focus groups and principal interviews supplied important direction that served as a source of validation while only receiving a formal check for face validity. The grounded theory analysis revealed the initial categories from the bottom-up approach, and the case study

methodology used a top-down approach to refine the categories. Ultimately the literature review added information necessary to design the NCEES Teacher Survey and NCEES Principal Interview questions (see Figure 1 in Chapter 1).

3. Following the derivation of construct and content validity for the NCEES Teacher Survey and the NCEES Principal Interview questions, both data collection instruments were checked for item validity. A committee of five teachers (three English teachers, one math teacher, and one history teacher) and one administrator (former English teacher) were given the respective data collection instruments and asked to verify item validity. Gay et al. (2009) defined item validity as the ability of the items of a data collection instrument to measure the content under study.

The teachers did not complete the questionnaire, and the administrator was not asked to provide answers to the interview questions. Also, the literature review was not supplied to any parties involved. The chosen teachers and administrator previewed the content for item validity and were asked for input on whether they understood the language and meaning of the questions. Specifically, the teachers and administrators were asked to comment on any language in any of the items on the instruments they did not understand and would have difficulty answering as a result. The teacher survey and principal interview questions were refined after about 2 weeks of reflection and input and after resending edited questions from the teacher survey back to participating teachers three times and to the principal twice.

4. The NCCES Teacher Survey and the NCEES Principal Interview were administered and completed by participating teachers and principals. Approval was received from the dissertation committee overseeing this study and the superintendent of the district under study. After approval, email addresses for teachers were confirmed

using data from the county and included any classroom teacher who received classroom observations and evaluation using NCEES during the 2014-2015 school year.

The NCEES Teacher Survey was administered online using Qualtrics. The window of completion time occurred within two time periods, each consisting of 2 weeks. The first window opened the moment after an initial email was sent out introducing the study. The initial email included a link to the Qualtrics questionnaire. After 2 weeks, a final notice was emailed to potential participants reminding them that there were 2 weeks left to respond. That email also contained a link to the Qualtrics online survey.

5. After data from the unstructured portion of the NCEES Teacher Survey were collected alongside data from the NCEES Principal Interview questions, grounded theory was used to generate theories as to the impact of NCEES on teaching practices and teacher leadership on teachers in the six high schools in the county under study. Statistical analysis of NCTSR also occurred during this time period.

6. The results were added to Chapter 4 of this study and analyzed in Chapter 5.

Anticipated Outcomes

The first anticipated outcome involved the quantity of classroom observations and evaluation received by teachers in the six high schools in the county under study. It was predicted that many teachers would see administrators as too busy to be involved in the classroom observation and evaluation processes at a level necessary to provide useable feedback. This was the sentiment in the recent Oakes and Robertson (2014) survey of North Carolina teachers; where of 795 respondents to the question, 52% believed the evaluator spent an adequate amount of time observing in order to make a valid evaluation, while 38% disagreed (10% were not sure). This result paralleled observations

by Bryant (2013) and the TNDOE (2012) study.

The expected results for the validity and reliability of NCEES including Standards 1-5 were mixed in the mind of the researcher. Given the pilot data were collected over 2 years previously and the focus group and principal interview questions for the pilot data were purposely open-ended, it was difficult to predict feedback from teachers and principals after the survey questions had been designed with more specificity. Also, the pilot data were small in sample, although many hours of dialogue were transcribed for the sample. NCEES had just been formally implemented during the collection of pilot data. This was one reason why teachers in the pilot data harped on the usability of NCEES being online. The online usability was a topic that dominated teacher talk during the focus groups; but during analysis, it was decided that this topic did not carry enough importance or relevance to explore in the literature review. With time, teachers have probably improved their ability to access their account online.

One important topic in the teacher focus groups within the pilot data was the consistent belief that teachers perceived their classroom observations and evaluation as valid, though not as reliable, but their responses to validity were passive and superficial with no exactness. There was one veteran teacher who appreciated the "talks" she had with administrators after being observed and evaluated but offered no specific details about the feedback from the administrator or how it influenced her teaching practices or teacher leadership. This was the consensus of teachers–they thought being observed and evaluated was helpful, but they identified no specific feedback that changed how they taught or led.

An important outcome of the teacher focus groups in the pilot data was that teachers felt estranged from the observation and evaluation process. Teachers gave the

processes involved in Standards 1-5 of NCEES and teacher evaluation in general lip service, but they also admitted that the standards on the NCEES rubric did not dictate how they planned lessons or carried them out. One teacher very bluntly said that regardless of what the NCEES rubric expected, his number one priority was the kids. This teacher identified that meeting the needs of his students was the driving force behind what he did in his classroom, not what NCEES expected.

The reliability of observer judgments was perceived negatively in teacher focus groups in the pilot data. Teachers viewed the results of classroom observations and evaluations they received in the past as differing depending upon which administrator they had worked with. Some administrators appeared more stringent than others, but this was generally the case when teachers talked about classroom observations and evaluation they had received in the past, not necessarily using NCEES.

Given these outcomes from the pilot data, the researcher expected more of the same in terms of validity–teachers would agree that Standards 1-5 of NCEES have been helpful alongside its rubric but would not be very specific about how or why. In terms of the reliability of Standards 1-5, the researcher expected an improvement in reliability. Administrators had received intensive, countywide training using the NCEES rubric since the time of collecting pilot data. It was believed that this would make a difference and that teachers would notice. Overall, the validity and reliability of NCEES was expected to be positive by teachers.

These expected outcomes of the validity and reliability of principal ratings for Standards 1-5 aligned with the Oakes and Robertson (2014) survey. In that survey, over two thirds agreed with their evaluation, but only half thought they received feedback that improved their teaching (Oakes & Robertson, 2014). Over 70% of the respondents

thought their most recent ratings on Standards 1-5 accurately reflected their abilities.

The researcher expected questions on the NCEES Teacher Survey to yield negative mean scores for questions pertaining to Standard 6 EVAAS®. This predicted outcome also aligned with the Oakes and Robertson (2014) survey where 49.29% of teachers (N=777) agreed and 12.74% strongly agreed with the following question: "I struggle with understanding how to change my practice through the use of Standard 6 data" (p. 10). When asked about how the use of Standard 6 would affect the morale of teachers in their building, 89% of teachers in the Oakes and Robertson survey predicted a negative impact (7.59% no impact, 3.54% positive impact, 0.25% strong positive impact). Also, 54% of teachers in the Oakes and Robertson survey predicted a negative impact on the quality of instruction based on the use of Standard 6.

In terms of using NCEES outcomes formatively and in a summative manner, the researcher hypothesized that teachers would respond positively proportional to years of experience. Teachers have historically criticized teacher evaluation systems for being perfunctory, paper-and-pencil exercises (Danielson, 2010/2011; Weisberg et al., 2009). With the exception of Standard 6 EVAAS® data, the researcher predicted that teachers would appreciate the fact that NCEES has made strides to make sure all teachers received feedback compared to the past where in many cases they were not observed or evaluated at all. Whether or not the feedback contained quality suggestions and insights that built teacher self-efficacy remained unpredictable. The researcher thought the answers to these predictions would become evident when comparing the perceptions of the Teacher+1 group to the Teacher+0 group across survey responses to both structured and unstructured items.

For the quantitative data (NCTSR), the researcher predicted that the distribution

of NCTSR would be skewed towards inflated results closely resembling Weisberg et al.'s (2009) analysis. At first glance, this appeared to be the case before running any statistical tests. The researcher had no predictions about the movement of NCTSR across years.

For the qualitative responses to the principal interviews, it was predicted that principals concentrate their discussions on the same thing the principals did in the pilot data: a lack of time. Although the TNDOE (2012) study outlined that TNDOE was paring back the number of classroom visits, no news of such an action in North Carolina has been noticed. However, North Carolina requires only two informal visits for veteran teachers compared to four in Tennessee. Still, the topic of time dominated the principal interviews in the pilot data.

Other than the topic of time, it was difficult for the researcher to predict how administrators would respond to the rest of the interview questions. There was little evidence visited in the literature about the perceptions of principals towards newly revamped teacher evaluation systems in terms of their validity or reliability. On the basis of Sartain et al. (2011), it was expected that principals would perceive Standards 1-5 as a valid measure of teaching practices and teacher leadership, but it was difficult to predict perceptions about reliability among principals. This was also the case with the perceptions of principals in terms of using VAM within teacher evaluation systems, although it seemed as if principals would appreciate another source of data describing teacher performance as a comparison alongside their judgments from classroom observations. Lastly, the researcher predicted that principals would have positive perceptions of the ability to offer formative feedback to teachers using Danielson's (2007) *Framework*. This prediction was based on the fact that the *Framework* was built on more than a decade of research of what constituted quality teaching practices and

teacher leadership (Danielson, 2007). There was no prediction by the researcher in terms of the perceptions of principals towards summative evaluation as the literature or pilot data offered little guidance.

Chapter 4: Results

Introduction

The aim of this study was to investigate the effects of NCEES on teaching practices and teacher leadership in a mostly rural county in the southwest piedmont region of North Carolina. The study has been designed using mixed methods (QUAL-Quan).

Chapter 1 presented evidence of a lack of quality in teaching practices and teacher leadership by historically examining teacher ratings and comparing them to student outcomes. In Chapter 2, evidence was also examined stemming from classroom observations and evaluation of teachers in their classrooms and how there was a lack of innovation in teaching practices and teacher leadership. Chapter 2 focused on using pilot data to generate conceptual categories of both historic and present-day processes within teacher evaluation systems that have shown to affect teaching practices and teacher leadership. A questionnaire was developed using conceptual categories generated by pilot data and refined by an in-depth analysis of the literature. Those conceptual categories included processes surrounding the quantity of classroom observations and evaluation, the quality of classroom observations and evaluation, the use of EVAAS® within NCEES, and how outcomes from NCEES were being utilized.

Chapter 3 presented the methodology of collecting data which involved using the validated questionnaire within a survey process for the teachers in the six high schools under study. Chapter 3 also laid out the plans to bring multiple sources of quantitative and qualitative data together in an effort to triangulate results.

The results of this study are presented in this chapter using the same flow presented in Chapter 3 sections entitled “Forms of Data Collection and Uses” and

"Anticipated Outcomes." Chapter 5 of this study follows this same flow for the sake of triangulating and generating theories using grounded theory.

First, results for each structured item were presented in frequencies and displayed in a table format. This table also included combining the lower and higher item options ("strongly disagree" + "disagree" and "strongly agree" + "agree") for each structured item to provide a general direction without considering the neutral "neither agree nor disagree." The combination of the lower and higher item options was presented in percentages.

Second, the results of the NCEES Teacher Survey were presented with descriptive data analysis including the mean, median, mode, and standard deviation for each structured item in order to describe the measures of central tendency and deviation from the mean. There has been debate in the literature about which descriptive measure of central tendency best suits Likert data and data on the ordinal scale (Cohen et al., 2007). To satisfy different views, the mean, median, and mode were calculated and displayed.

To calculate the measures of central tendency, values "1," "2," "3," "4," and "5" were anchored to item options "strongly disagree," "disagree," "neither agree nor disagree," "agree," and "strongly agree" in order to calculate the descriptive statistics. The results were displayed in a table.

Third, before choosing inferential statistical tests, data from the NCEES Teacher Survey were analyzed for normality. Tests for skewness and kurtosis were checked using Shaprio-Wilk's test alongside visually inspecting histograms, normal Q-Q plots, and box plots. The various methods of inspection showed the vast majority of data were not normally distributed. Most data distributions showed $p<0.05$ using the Shapiro-Wilk's test under the following hypotheses:

H_0=No ascertainable difference in the data distributions across groups due to skewness or kurtosis;

H_1=Ascertainable difference in the data distributions across groups due to skewness and kurtosis.

(α=0.05)

Because the majority of data were not normally distributed and because ordinal scale Likert data have been considered nonparametric by definition (Cohen et al., 2007), the nonparametric Mann-Whitney and Kruskal-Wallis tests were chosen to test the differences among groups. The Mann-Whitney test was designed to test the difference between two independent groups by ranking them and providing a Mann-Whitney test statistic "U" (Cohen et al., 2007). The Kruskal-Wallis test was designed to be the Mann-Whitney equivalent to test multiple independent groups by ranking them and providing an "Adjusted H" statistic.

Both the Whitney-Mann U and Adjusted H statistics have been used to test the probability that two or more groups were significantly different by calculating probability levels (Cohen et al., 2007). The chosen alpha level for this research was chosen at 5%, a level of confidence acceptable for this research endeavor (Cohen et al., 2007). A p value for each statistic ("U" and an "Adjusted H") was calculated for each of the groups listed in the demographic section of the NCEES Teacher Survey. The calculated p values were then compared to the chosen alpha level to test the significance of the difference among groups.

In summary, responses to the structured items on the NCEES Teacher Survey were disaggregated based upon the initial demographic items (years of experience, gender, ethnicity/race, Teacher+0 and Teacher+1, and Schools A, B, C, D, E, and F). The

Mann-Whitney U test was then used to detect whether there was a significant difference between the responses of teachers based on their gender or whether they had experienced a previous teacher evaluation system (Teacher+0/Teacher+1). The Kruskal-Wallis test was carried out to detect any potential differences between teacher years of experience and schools.

Inferential statistical tests were not run for teachers based on their ethnicity/race. A large majority (84) of teachers who responded to the NCEES Teacher Survey were White, whereas three Black teachers, two Hispanic teachers, and one Native American/ American Indian teacher responded. This breakdown invalidated an inferential statistical approach for this group.

The last step of the quantitative analysis involved NCTSR. NCTSR was presented for the county under study. NCTSR data for the county were graphed using bar graphs and compared against each of the three observed sources, Weisberg et al.'s (2009) "Two Districts With a Four Point Rating Scale" (see Appendix B), Curtis's (2011) "SY 2009-10 Teacher Impact Ratings" (see Appendix C), Lynn et al.'s (2013) "Number of Teachers in Each Rating Category by Standard" (see Appendix F), and the aggregate state NCTSR-obtained online from NCDPI. A chi-square goodness of fit analysis identified whether NCTSR from the most recent year of results for the county under study were significantly different from each of the three observed sources.

NCTSR was also used to investigate if there was movement in teacher ratings across years of the application of NCEES. For this analysis, a chi-square goodness of fit analysis was used to test whether a significant movement up or down the ratings has occurred across years. Applicable bar graphs were provided for each comparison across years. Each of the six schools in the county under study was analyzed.

A chi-square goodness of fit statistical test was chosen for both testing the NCTSR movements and distribution because of its ability to determine whether there was a difference in an observed sample of data compared to an expected sample. The χ^2 statistic was computed and used to calculate a probability that was compared to a preset alpha level of $\alpha=0.05$. This approach was taken to test if there was a significant difference across years of application of NCEES and whether there was a distribution of teacher ratings comparable to the three observed sources.

Lastly, the qualitative data from teachers and principals served were analyzed using grounded theory, and the results were presented in narrative form. The principals supplied over 6 hours of combined interview data, while teachers provided substantial qualitative data in the unstructured items of the NCEES Teacher Survey. The results of the grounded theory exploration of the qualitative data were presented in narrative form.

The qualitative data formed the backbone of this study, while the quantitative forms of data were used in an attempt to reinforce theories about how NCEES has affected teaching practices and teacher leadership in the six high schools in the county under study. By bringing together the various forms of data, results were triangulated and conclusions sought. The outcome of this final analysis is supplied in Chapter 5 in narrative form.

NCEES Teacher Survey Structured Item Analysis

The questionnaire was distributed online to a population of 299 high school teachers across six local high schools in the county under study. After 2 weeks, the researcher received results from 22 (23.4%) of the total N=94 respondents. A reminder was sent out after 2 weeks, and the researcher received results from the rest of the

sample, which included 72 (76.6%) responses. The total respondents were N=94; and in summary, 22 responded in the first 2-week window and 72 responded in the second 2-week window.

This yield was predictable. The first 2-week window opened as state testing had begun for students, and teachers were expectedly busy. However, it was imperative the survey process began after postconferences were finalized and teachers received year-end summative ratings. This was especially important for first-year teachers, because they would not have been able to participate in the survey process without having completed the full observation and evaluation cycle. This wait time forced the delay in beginning the survey process until the very end of the year, the busiest time of the year for everyone in a school building.

The second 2-week window began on the first workday (Monday) for teachers after the last day of school for students (Friday). State testing was completed, and teachers were in a better position to participate. The second 2-week window continued after the last workday for teachers and into summer break. The second 2-week window proved to be the most popular time of participation for teachers.

Of the 72 respondents in the second 2-week window, 44 respondents returned responses on the first day of the second 2-week window, which was the first teacher workday (Monday) after testing was completed. The final yield was 94 questionnaires received, although not every structured item was completed for all 94 questionnaires.

Qualtrics recorded 107 entries into the questionnaire and 94 responses: 80 were completed responses, 10 partial responses, and 4 responses were discarded because of a lack of data. This yielded a 26.8% response rate (80/299) or 30.1% (90/299) if the 10 partial responses were considered pooled alongside the completed responses. Also, the

completion rate was calculated at 84.1% (90/107) and associated dropout rate of 15.9% (17/107).

Qualtrics was programmed by the researcher to accept questionnaires as complete if the respondents were able to proceed through the questionnaire and click “complete” at the end. The goal in using this approach was to capture as much data as possible; and if it occurred that a respondent did not feel comfortable in providing a response to a certain item, the respondent could still provide valuable data for other structured and unstructured items. Respondents were directed to this approach and intention of the researcher within the Informed Consent section of the survey process.

Of the 10 partially completed questionnaires, six respondents failed to complete Item 6 which asked them in which school they taught. The 10 respondents providing partial responses provided input on various structured and unstructured items but failed to provide responses for every structured item. The results of these 10 questionnaires were accepted and placed into an EXCEL database to be used within EXCEL and moved to SPSS. The incomplete items for these 10 respondents were kept blank but skipped using functions and formulas in EXCEL and SPSS so as to not have the value of their answer calculated as a zero, in turn harming outcomes from statistical procedures.

Item 1 was the Informed Consent statement that was mandatory for respondents to answer in order to enter the questionnaire. If respondents failed to respond, they were disallowed to enter the questionnaire and were forced to either exit out of Qualtrics or be frozen on this particular item. Table 1 below shows the results for Item 1.

Table 1

Item 1: I have read and understood the consent form and desire of my own free will to participate in this study.

Response	Frequency (N=94)	Percentage
Yes	94	100%
No	0	0%

Item 2 asked teachers about their years of experience. Teachers with between 1 and 5 years of teaching experience participated close to double the rate compared to teachers with 6-10 years of teaching experience and similarly more compared to teachers with 11-15 years of teaching experience and teachers with 16-20 years of teaching experience. Teachers with more than 20 years of teaching experience were not well represented in the sample. However, the proportion of teachers based on their years of teaching followed the trend of teaching experience in North Carolina where newer teachers outnumber veteran teachers. In a 2012 study, the University of North Carolina (UNC, 2012) studied teacher recruitment and found that "The greening effect is evidenced by the change in years of teaching experience, from the modal years of experience in 1987-1988 at 15 years, to the modal years of experience 2007-2008 at 1 year" (p. 6).

Given the UNC (2012) evidence, it was a logical extension that teachers with less years of teaching experience would outnumber teachers with more years of teaching experience participating in the survey process. Table 2 displays the teaching experience levels of the respondents in the NCEES survey as collected within Item 2.

Table 2

Item 2: Please provide your years of teaching experience (private/public–including any other county, state, or country).

Years of Experience	Frequency (N=90)	Percentage
1-5	31	34.4%
6-10	17	18.8%
11-15	14	15.5%
16-20	16	17.8%
21-25	7	7.8%
26-30	3	3.3%
>30	2	2.2%

Item 3 asked teachers about their gender. Table 3 displays the results which show that female teachers responded at a rate of about three times higher than male teachers. This outcome roughly reflected the gender breakdown of high school teachers in North Carolina and across the U.S. at the time of this study.

Table 3

Item 3: Please provide your gender.

Gender	Frequency (N=90)	Percentage
Male	24	26.7%
Female	66	73.3%

Item 4 asked teachers about their ethnicity/race. Table 4 displays the results which

show that White teachers outnumbered others. Teachers with diverse ethnic/racial backgrounds were not well represented in this sample.

Table 4

Item 4: Please provide your ethnicity/race.

Ethnicity/Race	Frequency (N=90)	Percentage
Hispanic or Latino	2	2.2%
Black	3	3.3%
White	84	93.3%
Native American or American Indian	1	1.1%
Asian or Pacific Islander	0	0%

Item 5 asked teachers whether they had experienced classroom observations and evaluation using a teacher evaluation system prior to NCEES and was a pivotal item in the context and inferential statistical approach of this study. Teachers who had not received classroom observations and evaluation using a previous teacher evaluation system were labeled "Teacher+0" (pretest group), and teachers who had received classroom observations and evaluation using a previous teacher evaluation system were labeled "Teacher+1" (posttest group).

An important check for internal validity was carried out for Item 5. From the data in Table 2, every teacher with 11 or more years of experience (11-15, 16-20, 21-25, 25-30, >30) responded "Yes" to Item 5. This was critical because it would have been impossible for a teacher with 10 or more years of teaching experience to have not experienced a previous teacher evaluation system whether in North Carolina, another

state, or a private school. This meant that at least this subsample of respondents understood the meaning of Item 5.

It would have been optimal to have had the same capability of internally checking the responses of teachers with 1-5 or 6-10 years of experience, but these teachers were not subject to the same internal check because it would have been possible for these teachers to have experienced a previous teacher evaluation system in another state or a private school before they arrived in the county under study. It was also potentially possible for a sixth-year teacher to answer "no" to Item 5 as the NCEES rubric had rolled out in the county under study during the 2009-2010 school year, which was 6 years before the 2014-2015 school year.

Both of these scenarios are apparent in the data in Table 5. There were 10 teachers between 1-5 years of teaching experience that responded "Yes" to Item 5, meaning they had experienced a previous teacher evaluation system to NCEES. Four teachers with 6-10 years of teaching experience responded "No" to Item 5, which was taken to mean these particular teachers had been in the county under study since the 2009-2010 school year. It was also possible that teachers with 7 years of teaching experience could have answered "No" to Item 5 because it was possible that these teachers could have transferred into the county under study from a surrounding county in North Carolina that was piloting NCEES during the 2008-2009 school year. The latter case seemed unlikely, but there was no way to internally check this possibility.

The details surrounding the two groups of teachers with 1-5 and 6-10 years of teaching experience in no way invalidated their responses to the rest of the items on the questionnaire or the inferential statistical approach used to disaggregate their responses. The researcher found it valuable to note the impossibility of internally validating their

responses unlike the validation check for teachers with 11 or more years of teaching experience.

Pointing out response options of teachers with 1-5 and 6-10 years of experience also served to validate the differences in responses from teachers in Table 2 compared to Table 5. For example, this explanation sufficiently explained how it was possible that 31 teachers responded to having between 1-5 years of teaching experience in Table 2, yet only 25 responded "No" in Table 5 having never operated under a prior teacher evaluation system.

After investigating the raw data case by case and tallying results, it was observed that among the 25 teachers in Table 5 who answered "No" to having experienced a previous teacher evaluation system to NCEES, 21 teachers had 1-5 years of teaching experience and four had 6-10 years of teaching experience. Furthermore, there were 10 teachers with 1-5 years of teaching experience who received classroom observations and evaluation using a teacher evaluation system prior to NCEES and answered "Yes" to Item 5. Whether the teachers with 1-5 years of teaching who responded "Yes" entered the county under study from another state, country, or perhaps a private school was unknown and irrelevant for sake of this study. Table 5 displays the results for Item 5.

Table 5

Item 5: I have received classroom observations and evaluation using a teacher evaluation system prior to NCEES (e.g., under former North Carolina systems: TPAI [1990s], TPAI-R [2000s], or under any other system from other states or countries).

Response	Frequency (N=90)	Percentage
Yes (Teacher+1)	65	70.0%
No (Teacher+0)	25	30.0%

Item 6 asked teachers to identify the school where they taught. Six teachers failed to identify the school in which they taught after completing the first five demographic questions, although after skipping Item 6, these respondents did go on and complete other items on their questionnaires. The reason for this outcome was unknown other than perhaps these teachers did not want to divulge their place of employment regardless of anonymity being guaranteed throughout the invitation and Informed Consent Form. Because of the value of responses to other items, responses to these six questionnaires were retained in the database regardless of the incompleteness; however, these six responses were discarded when they disaggregated based upon each of the six schools in the county under study because of the inability to properly tabulate responses. Table 6 displays the results for Item 6 with letters substituted for names of the schools for the sake of anonymity.

Table 6

Item 6: I am a classroom teacher at the following school.

Schools	Frequency (N=84)	Percentage
A	19	22.6%
B	19	22.6%
C	31	36.9%
D	1	1.2%
E	12	14.3%
F	2	2.4%

Teacher responses to the NCEES Teacher Survey seemed to follow a pattern

noted by Cohen et al. (2007): Teacher responses centered on and around “neither agree nor disagree.” This has been an outcome observed in questionnaires because “most of us would not wish to be called extremists; we often prefer to appear like each other in many respects”; and, as a result, “this means we might wish to avoid the two extreme poles at each end of the continuum of the rating scales, reducing the number of positions in the scales to a choice of three (in a 5-point scale)” (Cohen et al., 2007, p. 327). Some items seemed to avoid the tendency to avoid the two extreme ratings, notably Items 7, 9, 11, 35, and 37. These results are displayed in Table 7, alongside the overall percent negative and positive.

Table 7

Frequency of Teacher Responses

NCEES Teacher Survey Item	N	SD	D	N	A	SA	%Neg	%Pos
7. Adequate Formal Observations	88	1	16	10	41	20	19	69
9. Adequate Informal Observations	88	4	20	15	34	15	27	56
11. Adequate Time	89	3	19	14	39	14	25	60
13. Ratings for Standards 1-5 align with 6	65*	4	19	21	19	2	35	32
15. Appropriateness of Standards 2-5	85	1	25	11	44	4	31	57
17. Appropriateness of Standard 1	85	3	16	20	43	3	22	54
19. Validity of NCEES–Teaching Practices	86	3	23	13	42	5	30	55
21. Validity of NCEES–Leadership	87	5	23	22	34	3	32	43
23. Reliability of Ratings–Standards 1-5	77*	3	15	14	39	6	23	58
25. Reliability of Observers	77*	3	16	22	31	5	25	47
27. Validity of EVAAS®–Student Data	68*	8	22	18	17	3	44	29
29. Validity of EVAAS®–Teacher Data	68*	9	20	21	14	4	43	27
31. Reliability of EVAAS®	64*	5	19	15	19	6	38	39
33. Unintended Consequences–EVAAS®	64*	7	17	13	23	4	38	42
35. Formative Feedback–Practices	82	4	30	22	16	10	42	32
37. Formative Feedback–Leadership	81	5	10	23	33	10	19	53
39. Summative Feedback–Practices	80	5	31	23	18	3	45	26
41. Summative Feedback–Leadership	80	9	10	31	24	6	24	38

Note: N=sample size, SD=strongly disagree, D=disagree, N=neutral, A=agree, and SA=strongly agree;
**Excludes teachers not receiving EVAAS® scores or who had not received enough data to form an opinion.*

Overall, five items were found to elicit negative responses from teachers–Items 13, 27, 29, 35, and 39. Items 13, 27, and 29 involved perceptions centered on Standard 6 and EVAAS® outcomes, and Items 35 and 39 involved perceptions centered on feedback. In instances where teacher responses were overall negative, the proportion was not

overwhelming. Item 39 received the largest proportion of negative to positive responses at a rate approaching a 2:1 ratio (1.71).

The positive responses from teachers had high a proportion of positive-to-negative perceptions with the exception of Items 21, 33, and 41. The rest of the positive responses from teachers either approached or surpassed a 2:1 ratio. With the exception of Items 21, 33, and 41 the remainder of positive-to-negative ratios were greater than even the highest of the negative-to-positive ratios of which was 1.73 for Item 39.

Positive-to-negative and negative-to-positive ratios were calculated by dividing the percent positive or negative by the converse with 1.00 being an exact 1:1 relationship or a “neutral” outcome. The exact ratios of percent positive-to-negative or vice versa are observed in Table 8.

Table 8

Ratios of Positive-to-Negative or Negative-to-Positive Responses

Item	%Negative-to-Positive Ratio	%Positive-to-Negative Ratio
7. Adequate Formal Observations	-	3.63
9. Adequate Informal Observations	-	2.07
11. Adequate Time	-	2.40
13. Ratings for Standards 1-5 align with 6	1.09	-
15. Appropriateness of Standards 2-5	-	1.83
17. Appropriateness of Standard 1	-	2.46
19. Validity of NCEES–Teaching Practices	-	1.83
21. Validity of NCEES–Leadership	-	1.34
23. Reliability of Ratings–Standards 1-5	-	2.52
25. Reliability of Observers	-	1.88
27. Validity of EVAAS®–Student Data	1.52	-
29. Validity of EVAAS®–Teacher Data	1.59	-
31. Reliability of EVAAS®	1.03	-
33. Unintended Consequences–EVAAS®	-	1.11
35. Formative Feedback–Practices	1.31	-
37. Formative Feedback–Leadership	-	2.79
39. Summative Feedback–Practices	1.73	-
41. Summative Feedback–Leadership	-	1.58

The next step in the quantitative analysis of the NCEES Teacher Survey data was

to analyze the measures of central tendency. Because of the dispute of which measure of central tendency to use with ordinal, Likert scale data (Cohen, et al., 2007), the mean, median, mode, and standard deviation were calculated for the sake of satisfying potential conflicts. The mean was the average of the Likert scores, anchored at values of 1 for "strongly disagree," 2 for "disagree," 3 for "neither agree nor disagree," 4 for "agree," and 5 for "strongly agree." The median was the center value of the Items scores after aligning them from the lowest to the highest, and the modal value was the Item response chosen the most often. The standard deviation described the dispersal of scores around the mean (Cohen et al., 2007).

Table 9

Measures of Central Tendency

Item	Mean	Median	Mode	Standard Deviation
7. Adequate Formal Observations	3.72	4	4	1.05
9. Adequate Informal Observations	3.41	4	4	1.15
11. Adequate Time	3.47	4	4	1.09
13. Ratings for Standards 1-5 align with 6	2.94	3	3	0.98
15. Appropriateness of Standards 2-5	3.29	4	4	0.99
17. Appropriateness of Standard 1	3.32	4	4	0.94
19. Validity of NCEES–Teaching Practices	3.27	4	4	1.03
21. Validity of NCEES–Leadership	3.08	3	4	1.01
23. Reliability of Ratings–Standards 1-5	3.39	4	4	1.02
25. Reliability of Observers	3.25	3	4	0.99
27. Validity of EVAAS®–Student Data	2.78	3	2	1.09
29. Validity of EVAAS®–Teacher Data	2.76	3	3	1.11
31. Reliability of EVAAS®	3.03	3	2	1.14
33. Unintended Consequences–EVAAS®	3.00	3	4	1.15
35. Formative Feedback–Practices	2.98	3	2	1.12
37. Formative Feedback–Leadership	3.41	4	4	1.06
39. Summative Feedback–Practices	2.79	3	2	0.99
41. Summative Feedback–Leadership	3.10	3	3	1.09

In analyzing the measures of central tendency, the mean values for items on the

NCEES Teacher Survey resulted in outcomes similar to the analysis in Table 8 of the negative-to-positive and positive-to-negative ratios. Items 13, 27, 29, 31, 35, and 39 received the lowest mean scores, which closely reflected the negative-to-positive ratios in Table 8. Items with high positive-to-negative ratios in Table 8 experienced higher mean outcomes reflected in Items 7, 11, 17, 23, and 37.

Both the median and modal values also paralleled outcomes in Table 8. Items 13, 27, 29, 31, 35, and 39 received the highest negative-to-positive ratios, and these same items all had lower medians of a 3. Likewise, Items 13, 27, 29, 31, 35, and 39 all received lower modal values of 3, 2, 3, 2, 2, and 2, respectively. Items with high positive-to-negative ratios experienced higher modal values of 4 (Items 7, 11, 17, 23, 37).

To test the difference between groups, nonparametric tests were used. The Kruskal-Wallis test was used for the years of experience of teachers and each school. The Mann-Whitney test was used for gender and the Teacher+0 and Teacher+1 groups. The results were displayed in Table 7 indicating the U-value statistic for the Mann-Whitney test and the Adjusted H value for the Kruskal-Wallis test, alongside the p value when the alpha level was set at 5%.

Table 10

Mann-Whitney and Kruskal-Wallis Statistics

Item	Group	N	U/H	p
7. Adequate Formal Observations	Yrs Exp	88	H=4.591	0.597
	Gender	88	U=618.500	0.135
	+0/+1	88	U=758.000	0.920
	School	88	H=6.087	0.298
9. Adequate Informal Observations	Yrs Exp	88	H=1.307	0.971
	Gender	88	U=688.000	0.556
	+0/+1	88	U=762.500	0.957
	School	88	H=4.119	0.532
11. Adequate Time	Yrs Exp	89	H=19.574	0.003**
	Gender	89	U=648.000	0.272
	+0/+1	89	U=662.500	0.252
	School	89	H=2.707	0.745
13. Ratings for Standards 1-5 align with 6	Yrs Exp	65*	H=7.351	0.290
	Gender	65*	U=355.500	0.301
	+0/+1	65*	U=307.000	0.595
	School	65*	H=3.524	0.474
15. Appropriateness of Standards 2-5	Yrs Exp	85	H=6.090	0.413
	Gender	85	U=673.000	0.826
	+0/+1	85	U=688.000	0.787
	School	85	H=9.430	0.093
17. Appropriateness of Standard 1	Yrs Exp	85	H=8.361	0.213
	Gender	85	U=679.000	0.879
	+0/+1	85	U=690.500	0.809
	School	85	H=5.860	0.320
19. Validity of NCEES–Teaching Practices	Yrs Exp	86	H=3.706	0.716
	Gender	86	U=669.000	0.709
	+0/+1	86	U=565.500	0.140
	School	86	H=2.325	0.803
21. Validity of NCEES–Leadership	Yrs Exp	87	H=3.704	0.717
	Gender	87	U=712.000	0.661
	+0/+1	87	U=657.500	0.247
	School	87	H=6.111	0.296
23. Reliability of Ratings–Standards 1-5	Yrs Exp	77*	H=2.724	0.843
	Gender	77*	U=506.000	0.422
	+0/+1	77*	U=464.500	0.269
	School	77*	H=2.317	0.678

(continued)

Item	Group	N	U/H	p
25. Reliability of Observers	Yrs Exp	77*	H=14.118	0.028**
	Gender	77*	U=542.500	0.584
	+0/+1	77*	U=433.500	0.063
	School	77*	H=7.786	0.100
27. Validity of EVAAS®–Student Data	Yrs Exp	68*	H=13.133	0.041**
	Gender	68*	U=365.500	0.157
	+0/+1	68*	U=266.000	0.025**
	School	68*	H=1.760	0.780
29. Validity of EVAAS®–Teacher Data	Yrs Exp	68*	H=12.031	0.061
	Gender	68*	U=438.500	0.869
	+0/+1	68*	U=296.000	0.120
	School	68*	H=1.904	0.753
31. Reliability of EVAAS®	Yrs Exp	64*	H=11.502	0.074
	Gender	64*	U=394.500	0.616
	+0/+1	64*	U=258.500	0.124
	School	64*	H=4.135	0.388
33. Unintended Consequences–EVAAS®	Yrs Exp	64*	H=10.130	0.119
	Gender	64*	U=300.500	0.403
	+0/+1	64*	U=762.500	0.957
	School	64*	H=0.748	0.945
35. Formative Feedback–Practices	Yrs Exp	82	H=14.772	0.022**
	Gender	82	U=573.000	0.343
	+0/+1	82	U=463.000	0.032**
	School	82	H=2.957	0.707
37. Formative Feedback–Leadership	Yrs Exp	81	H=11.331	0.079
	Gender	81	U=641.500	0.933
	+0/+1	81	U=462.500	0.058
	School	81	H=5.759	0.330
39. Summative Feedback–Practices	Yrs Exp	80	H=6.768	0.343
	Gender	80	U=547.000	0.303
	+0/+1	80	U=553.000	0.583
	School	80	H=6.266	0.281
41. Summative Feedback–Leadership	Yrs Exp	80	H=4.995	0.544
	Gender	80	U=540.000	0.269
	+0/+1	80	U=491.000	0.205
	School	80	H=2.734	0.741

*Note: N=sample size, U=Mann-Whitney test statistic, H=Kruskal-Wallis test statistic; Yrs Exp=years of experience, +0/+1=Teacher+0/Teacher+1 groups; *excludes teachers not receiving EVAAS® scores or who had not received enough data to form an opinion; **denotes significant difference at α=0.05.*

The Kruskal-Wallis test found a significant difference among groups in Items 11,

25, 27, and 35 for years of experience, while the Mann-Whitney test found a significant difference among groups in Items 27 and 35. For these results, the null hypothesis was rejected and alternative accepted because the probability level was calculated below the preset α=0.05 level. In Chapter 5, these results are triangulated in an attempt to uncover potential effects that these calculated differences had on teaching practices and teacher leadership. The triangulation efforts used qualitative data from principal interviews and teacher responses to the unstructured responses on the NCEES Teacher Survey.

In Item 25, the Mann-Whitney test result approached significant levels at 0.063; while for Items 29 and 31, the Kruskal-Wallis test approach significant levels at 0.061 and 0.074. For Item 37, both the Kruska-Wallis and Mann-Whitney tests produced p values approaching significant levels at 0.079 and 0.058.

For the significant differences among groups on Table 10, further graphical analysis attempted to uncover where the differences were located across groups. A series of graphs follow that resulted from this approach.

For Item 11, teachers with less years of experience seemed to agree at a higher rate proportionally that they had enough time to participate in pre and postconferences in using NCEES. In an attempt to determine the differences between groups, the neutral "neither agree nor disagree" response was excluded in the graphical analysis. As a result, the number of responses analyzed decreased from N=89 to N=75. The breakdown of Item 11 for teacher years of experience is displayed in Figure 3, noting that the number of teachers with more years of experience responding to the NCEES Teacher Survey was low reducing the ability to infer the differences among groups.

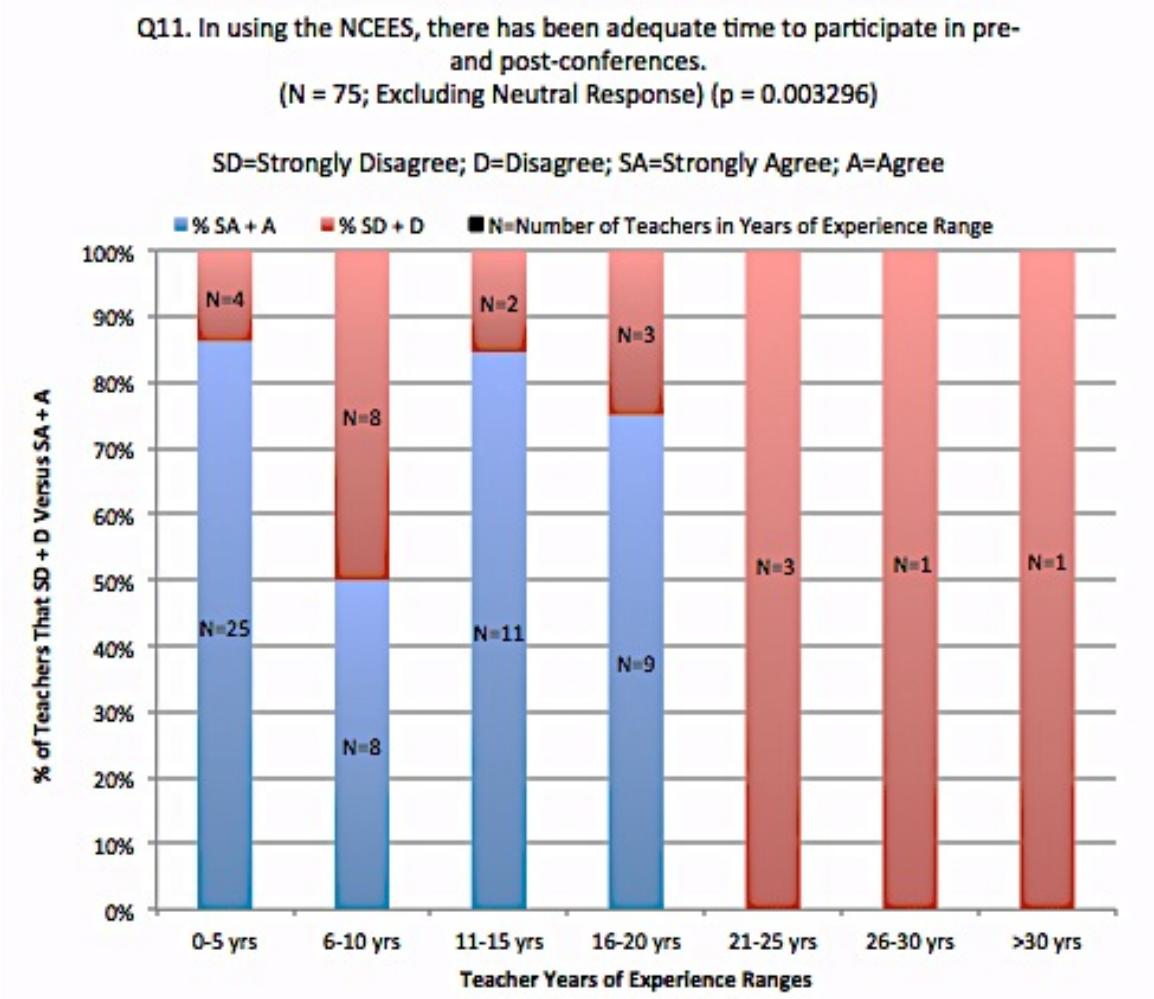


Figure 3. Responses to Item 11 across Teacher Years of Experience.

For Item 25, teachers with less years of experience seemed to agree at a higher rate proportionally that various observers showed agreement in rating their performance for Standards 1-5. In an attempt to determine the differences between groups, the neutral "neither agree nor disagree" response was excluded in the graphical analysis. As a result, the number of responses analyzed decreased from N=77 to N=55. The breakdown of Item 25 for teacher years of experience is displayed in Figure 4, noting that the number of teachers with more years of experience responding to the NCEES Teacher Survey was low reducing the ability to infer the differences among groups.

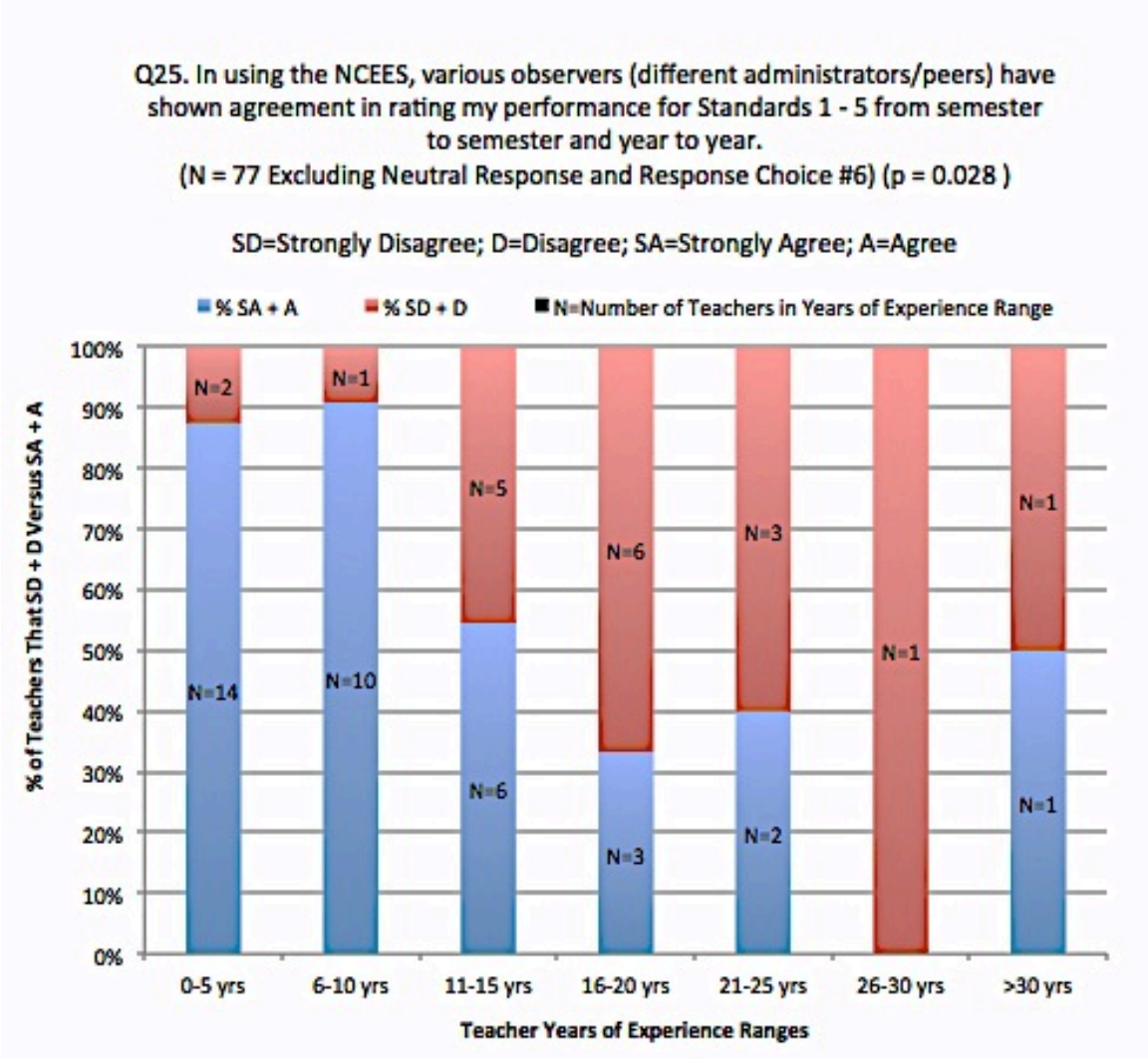


Figure 4. Responses to Item 25 across Teacher Years of Experience.

For Item 27, teachers with more years of experience seemed to agree at a higher rate proportionally that Standard 6 ratings reflected the academic growth of their students. In an attempt to determine the differences between groups, the neutral "neither agree nor disagree" response was excluded in the graphical analysis. As a result, the number of responses analyzed decreased from N=68 to N=50. The breakdown of Item 27 for teacher years of experience is displayed in Figure 5, noting that the number of teachers with more years of experience responding to the NCEES Teacher Survey was

low reducing the ability to infer the differences among groups.

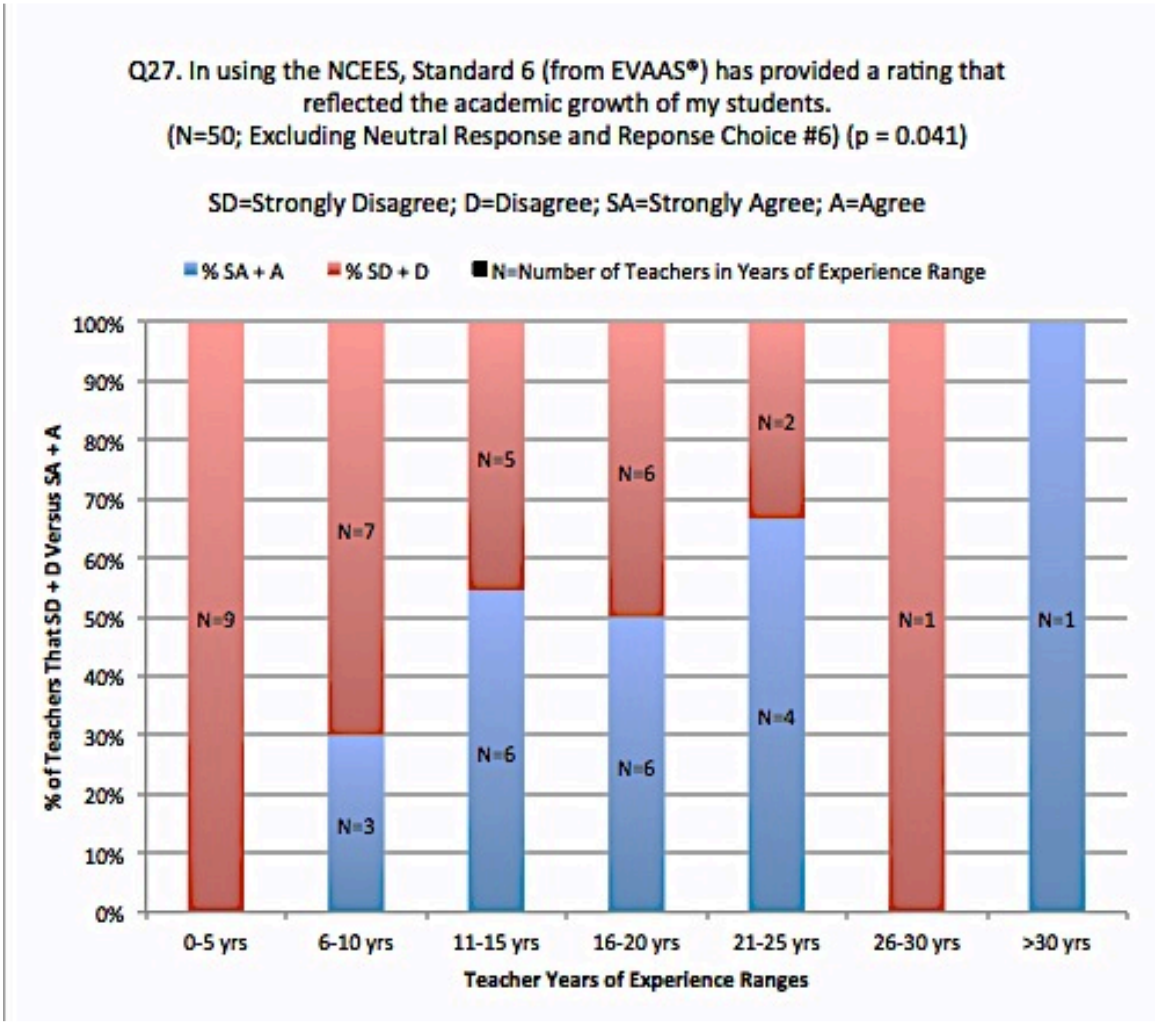


Figure 5. Responses to Item 27 across Teacher Years of Experience.

There was also a significant difference for Item 27 found across the pre and posttest groups Teacher+0 and Teacher+1. Teachers who had experienced a teacher evaluation system prior to NCEES seemed to agree at a higher rate proportionally that their Standard 6 rating reflected the academic growth of their students. The breakdown of Item 27 for Teacher+0 and Teacher+1 is displayed in Figure 6, noting that the neutral response "neither agree nor disagree" was excluded in the graphical analysis reducing N=68 to N=50.

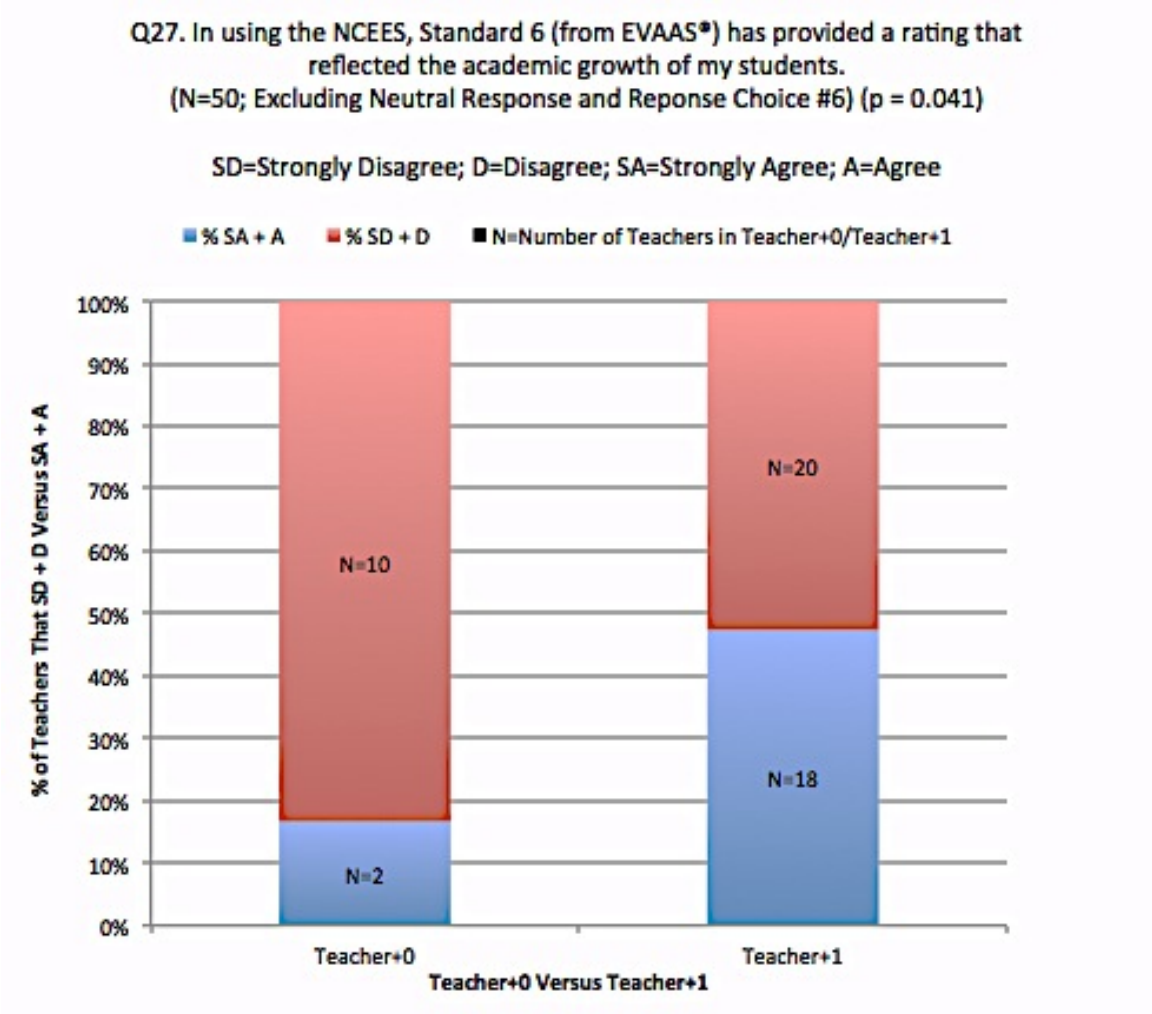


Figure 6. Responses to Item 27 across Teacher+0 and Teacher+1.

For Item 35, teachers with less years of experience seemed to agree at a higher rate proportionally that they received useful feedback from observers to improve their teaching practices. In an attempt to determine the differences between groups, the neutral "neither agree nor disagree" response was excluded in the graphical analysis. As a result, the number of responses analyzed decreased from N=82 to N=60. The breakdown of Item 35 for teacher years of experience is displayed in Figure 7, noting that the number of teachers with more years of experience responding to the NCEES Teacher Survey was low reducing the ability to infer the differences among groups.

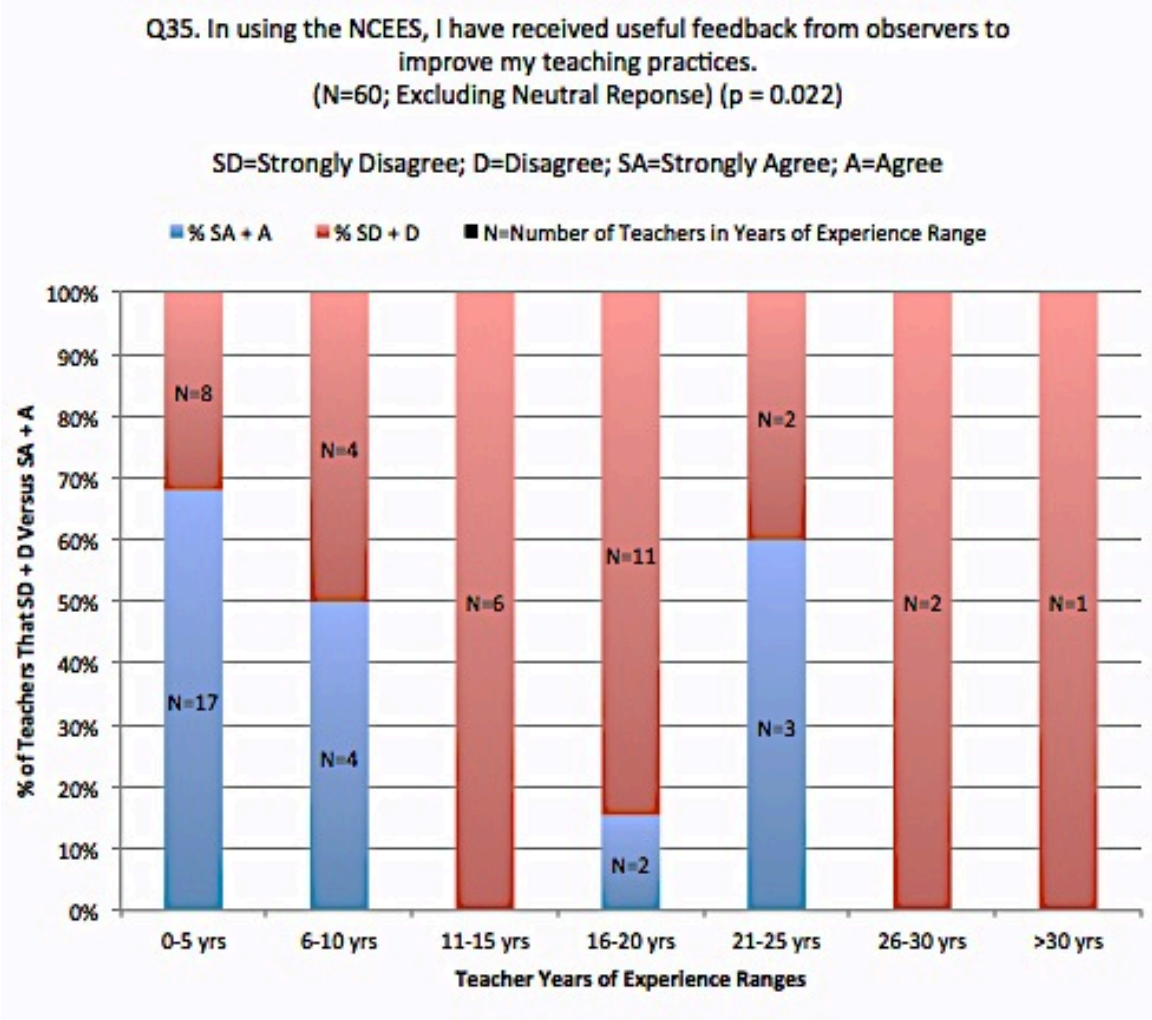


Figure 7. Responses to Item 35 across Teacher Years of Experience.

There was also a significant difference for Item 35 found across the pre and posttest groups Teacher+0 and Teacher+1. Teachers who had not experienced a teacher evaluation system prior to NCEES seemed to agree at a higher rate proportionally that they received useful feedback from observers to improve their teaching practices. The breakdown of Item 35 for Teacher+0 and Teacher+1 is displayed in Figure 8, noting that the neutral response "neither agree nor disagree" was excluded in the graphical analysis reducing N=82 to N=60.

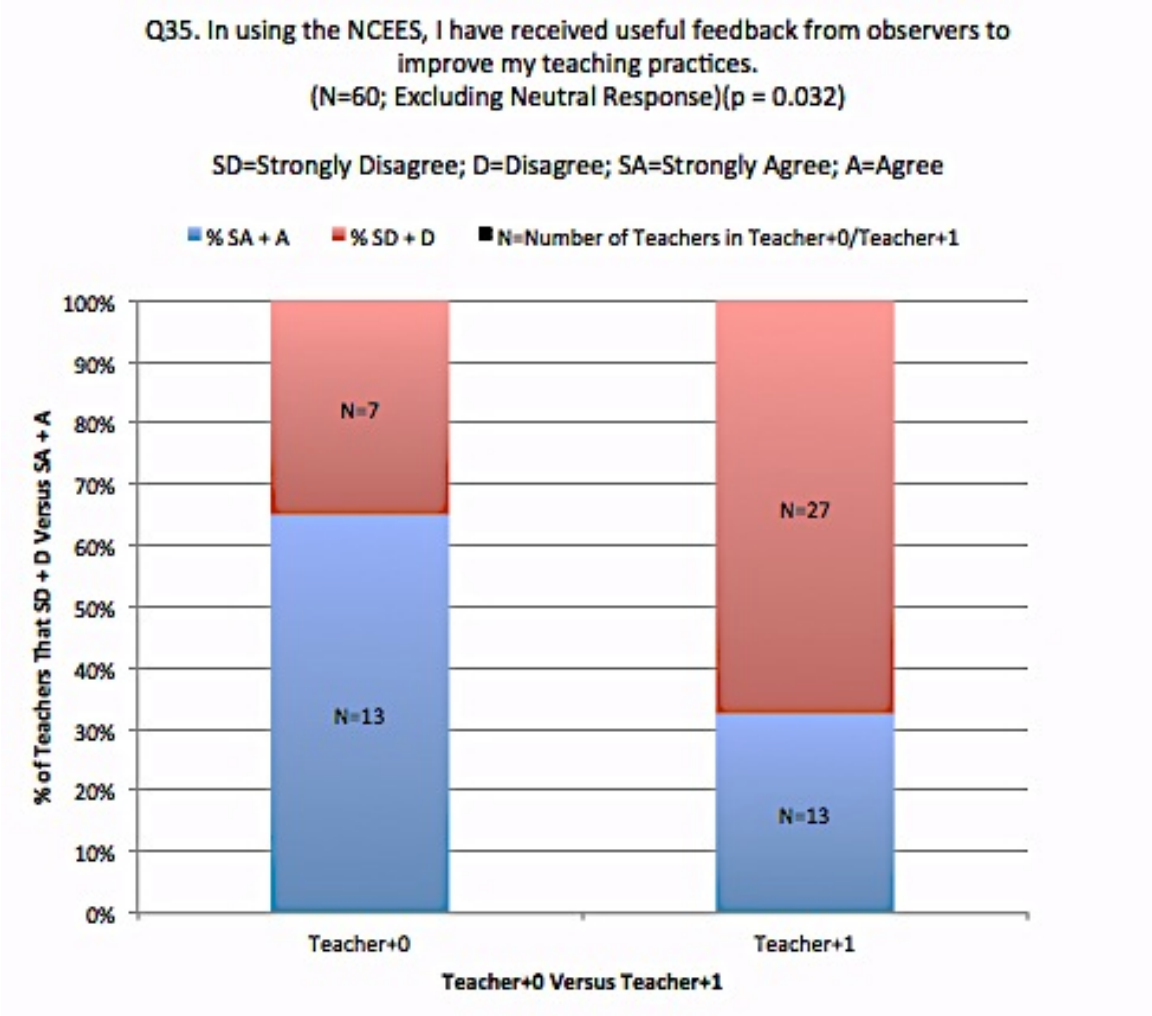


Figure 8. Responses to Item 35 across Teacher+0 and Teacher+1.

In summary, significant differences were found across groups for Items 11, 27, and 35 on the NCEES Teacher Survey. Item 11 was significantly different across years of teacher experience, and Items 27 and 35 were significantly different across years of teacher experience and the pre and posttest groups Teacher+0 & Teacher+1.

Responses across the demographic groups of gender and school revealed no significant differences among groups. There were not enough responses from Black teachers to test for differences.

After further graphical investigation, it was determined in Item 11 that teachers

with less years of experience agreed in higher proportion they had an adequate amount of time to participate in pre and postconferences while teachers with more years of experience disagreed. Teachers with more years of experience and who had experienced a previous teacher evaluation system prior to NCEES agreed in higher proportion in Item 27 that Standard 6 EVAAS® outcomes provided a teacher rating that reflected the academic growth of their students. Finally, teachers with less years of experience and who had not experienced a previous teacher evaluation system prior to NCEES agreed in higher proportion that they received useful feedback from observers to improve their teaching practices.

It is important to note that because of the limited number of teachers with more years of experience responding to the NCEES Teacher Survey, inferences should be used cautiously within the statistical outcomes. It was possible that with a limited number of teachers with more years of experience responding, small deviation in Groups 21-25 years, 26-30 years, and >30 years could have caused the differences across groups to be overstated for that demographic.

NCTSR Analysis of Teacher Ratings across Standards and Years

In order to ascertain the effects of NCEES on teaching practices and teacher leadership, a quantitative approach was taken to investigate movement of teacher ratings over time at the school and county levels. The goal was to determine whether teacher ratings had changed on the aggregate in response to improvements in teaching practices and teacher leadership due to the implementation of NCEES. Related to the NCEES, McREL et al. (2012) stated, “The instruments are designed to promote effective leadership, quality teaching, and student learning while enhancing professional practice and leading to improved instruction” (p. 4).

To detect movement of teaching ratings on the aggregate over time, teacher ratings from the 2011-2012 school year were compared to the last year of published teacher ratings from NCDPI's database (http://www.ncpublicschools.org/effectiveness-model/data/), which was the 2013-2014 school year. A chi-square statistic was calculated and compared to a preset alpha level of 0.05 to test if there was significant movement of teacher ratings. A later analysis would attempt to establish in what direction the ratings had moved.

Although NCEES was formally implemented in the 2010-2011 school year, newer teachers and renewing veteran teachers were overrepresented (Lynn et al., 2013). The 2011-2012 school year was the first year that all teachers were observed and evaluated across the state which provided a fair baseline of data (Lynn et al., 2013).

The 2011-2012 school year served as the "expected" value range and the 2013-2014 school year as the "observed" value range. This was in accordance with the null and alternative hypotheses presented in Chapter 3 where the null stated there would be no significant difference in NCTSR over time due to the implementation of NCEES. Based on the null hypothesis, ratings would have to change from the 2011-2012 year to the 2013-2014 school year, which logically made the 2011-2012 school year the start point and "expected" values within the analysis.

Cohen et al. (2007) advised that contingency tables for a chi-square analysis should contain cells where at least 80% have five or more cases, otherwise caution should be used in the analysis. To satisfy Cohen et al.'s prerequisite, rating categories across contingency tables were removed if they contributed to crossing the 80% threshold. In many cases, this usually removed data from one or the other of the "developing" and "distinguished" categories because they had zero ratings or would have contributed to

crossing the 80% threshold. In some cases, data from both "developing" and "distinguished" categories were removed because they contained zero ratings or would have contributed to crossing the 80% threshold. This explained the different degrees of freedom used across the analysis.

Standards 1 and 4, alongside Standard 6, were analyzed for NCSTR across years because Standards 1 and 4 represent the two areas inherent in this study. Standard 1 has been used in NCEES to gauge teacher leadership, and Standard 4 has been used to judge the quality of teaching practices during classroom observations and subsequent evaluation. Standards 1 and 4 were also used in an attempt to preserve higher sample sizes because all teachers have received classroom observations and evaluation using these two standards. The other standards in NCEES have only been used for newer teachers or veteran teachers who were renewing their teaching license; otherwise, veteran teachers have received an abbreviated observation and evaluation schedule where they were rated based on Standards 1, 4, and 6. This followed an approach by Batten (2013) that examined the correlation of Standard 6 EVAAS® outcomes with Standard 4 ratings.

Standard 6 was also assessed for NCTSR movement across the school years available. Only school years 2012-2013 and 2013-2013 were recorded in NCDPI's database; and as a result, the analysis centered on movement between these two school years for Standard 6. Standard 6 had lower sample sizes because not all teachers received EVAAS® outcomes.

It was important to point out that the number of teachers receiving ratings changed over the time period under study. The largest change was at School B where during the 2011-2012 school year, there were 73 teachers who received ratings compared to 63 during the 2013-2014 school year. The other schools encountered minor changes in

the number of teachers who received ratings within this time period with School A changing by two, School C by one, and School E by four. To account for the different sizes of the groups, the proportion of teachers in each rating category during the 2011-2012 school year was used to standardize the 2013-2014 data by recalculating the distribution within the 2013-2014 school year based on the proportion of ratings from the 2011-2012 school year. After being standardized, the resulting population of teachers under analysis was based on the number of teachers rated during the 2013-2014 school year and used as the "expected" data in the chi-square analysis.

Two schools (Schools D and F) did not have enough teachers to provide an analysis. One of these schools was a magnet school, and the other an alternative school; and both schools had less than 20 teachers. The results of the chi-square tests across years and standards are supplied in Table 11 for Schools A, B, C, and E and included the standardized sample sizes across years, the χ^2 statistic, degrees of freedom, and the calculated p value.

Table 11

Chi-Square Analysis of NCTSR Movement across Years and Standards 1 and 4

School	Standard	N	χ^2	df	p
A	1 Leadership	75	21.33333	2	0.0000***
	4 Practices	75	17.30640	2	0.0001***
	6 EVAAS®	43	1.89610	2	0.3875
B	1 Leadership	63	9.05034	1	0.0026**
	4 Practices	63	4.88470	1	0.0271*
	6 EVAAS®	33	4.81558	2	0.0900
C	1 Leadership	80	4.12019	2	0.1274
	4 Practices	80	14.74253	1	0.0001***
	6 EVAAS®	47	13.50763	2	0.0012**
E	1 Leadership	69	1.26359	2	0.4920
	4 Practices	69	1.20478	1	0.2724
	6 EVAAS®	42	2.08125	2	0.0353*

*Note. *denotes significant difference at* α*=0.05; **denotes significant difference at* α*=0.01; ***denotes significant difference at* α*=0.001.*

It was determined by the chi-square analysis that School E was the only school that had stable ratings across years, and the ratings were not significantly different as a result. However, School E had significant changes in its Standard 6 ratings across the 2 years of available data. School C was stable across years for Standard 1 only. The null hypothesis was rejected in all but three cases after observing the significant differences in teacher ratings across years due after the implementation and application of NCEES. The null hypothesis was rejected in two cases (Schools C and E) and accepted in two others (Schools A and B) for the rating movement across years in Standard 6.

In an attempt to discover the direction of NCTSR movement across years, an analysis followed which tracked ratings by compiling them by rating category and

movement across years. The basis for the highly significant differences observed in Table 11 was manifested within this analysis. The results of the compilation of NCTSR movement across years is displayed in Table 12, and the results ultimately showed the net gain and loss among rating categories between the 2011-2012 and 2013-2014 school years.

Table 12

NCTSR Movement across Years by All Rating Categories and All Standards

School Year	Rating Category	Rating Frequency	Net Loss/Gain
2011-2012	Distinguished	105	-
	Accomplished	573	-
	Proficient	371	-
	Developing	19	-
2013-2014	Distinguished	30	-75
	Accomplished	542	-31
	Proficient	483	+112
	Developing	8	-11

The net movement of ratings towards the "proficient" rating provided the basis for the highly significant differences between years apparent in Table 11. However, it should be noted that movements for specific teachers up and down the rating categories were not tracked in this analysis. There was also no attempt to track new teachers hired within this time period.

The results of this analysis placed 3% of teachers as "distinguished," 51% of teachers as "accomplished," 45% of teachers as "proficient," and 1% of teachers as "developing" across all standards on NCEES. It is important to note that these totals were compiled across all standards for both school years under analysis, and that one school

under study was flagged by NCDPI's database as indicating less than five teachers were evaluated. The results for this small school with a small number of teachers were not included, meaning there could be a +/- error of five ratings for Standards 2, 3, and 5; however, this would not affect the outcome of this analysis by a meaningful extent.

Teachers noted the tendency of being rated proficient in unstructured items dealing with Standards 1-5 on the NCEES Teacher Survey. Teachers responded with comments such as

- I feel that it is about impossible to be rated above proficient because it is really hard to show and prove.
- It has also created the ideal that the bare minimum (Proficient) is ok to keep one's job.
- It is confusing to hear different things from different people and get different ratings, although at the end it has seemed as though they pan out to me being "proficient" anyways.

The goal in this analysis was to show the aggregate movement of the NCTSR over school years to detect whether there was any movement signifying improvements in teaching practices and teacher leadership. It is also important to note that just because there was an aggregate movement towards the "proficient" rating category does not necessarily mean that teacher performance was lacking in quality. On the contrary, the qualitative data supplied by principals indicated that with training over the years, they have become more accurate raters of teaching practices and teacher leadership. This topic is explored further in Chapter 5.

NCTSR Analysis of Teacher Ratings across Three Observed Sources

NCTSR was also used for the schools in the county under study to compare against three observed sources: Weisberg et al. (2009), Curtis (2011), and Lynn et al. (2013). The three observed sources were chosen because of their status in the literature.

The null hypotheses for all three observed sources focused on finding whether there was a difference in NCTSR for the schools under study compared to the three observed sources without indicating whether the teacher ratings for the schools were perhaps "better" or "worse" or more like one observed source or others. After finding out whether there was a significant difference in ratings for the schools under study compared to the three observed sources, an investigation ensued to find out which of the three sources the teacher ratings best reflected. This approach was taken to guard against interjecting bias that NCSTR for the schools under study would appear one way or another–the focus was on investigating whether there was a *difference* and then investigating further.

Weisberg et al.'s (2009) distribution (see Appendix B) of teacher ratings across various school districts has served as a primary example in the literature of how teacher evaluation systems have not historically differentiated among teachers based upon performance–virtually all teachers have historically been found good or great. The goal in comparing the distribution of ratings under NCEES to Weisberg et al.'s distribution was to explore whether NCEES has been successful in differentiating among teachers based on performance in such a way that ratings have been more balanced across categories. It would also be important to compare this outcome to qualitative data supplied by principals during their interviews and teachers on the NCEES Teacher Survey in order to observe how potential shifts in the distribution of teacher ratings has effected teaching

practices and teacher leadership. The null hypothesis from Chapter 3 stated there would not be a significant difference between Weisberg et al.'s distribution of teacher ratings and those of NCEES.

Curtis's (2011) distribution (see Appendix C) of teacher ratings in the D.C. public schools demonstrated an improvement in ratings across four rating categories compared to that of school districts with four rating categories in Weisberg et al.'s (2009) analysis. D.C.'s teacher evaluation system has rating categories similar to NCEES in structure and definition. The D.C. teacher evaluation systems modeled their instrument using Danielson's (2007) *Framework* to rate teacher performance, the same instrument used within NCEES. Most importantly, Curtis's distribution was chosen because it modeled the changes set forth in the RttT initiative in that teacher ratings moved away from Weisberg et al.'s distribution observed during the 2008-2009 school year in the D.C. public schools (see Appendix C–bottom), to a more balanced distribution (see Appendix C–top). The overall goal in comparing NCTSR to Curtis's findings was to explore whether NCEES has similarly distributed teacher ratings across categories and looking to the qualitative data to detect how such a potential shift has affected teaching practices and teacher leadership.

The schools in the county under study were also compared with findings from Lynn et al. (2013). The goal in this analysis was to observe whether the ratings in the schools under study followed the clustering of ratings around the "proficient" and "accomplished" ratings as was found to be the case in the sample of North Carolina teachers in Lynn et al. Table 12 showed the net gain/loss of ratings across categories. The outcomes in Table 12 provided insight that the schools under study would perhaps reflect the same clustering reflected in Lynn et al.

As with the analysis of the Weisberg et al. (2009) and Curtis (2011) studies, the results from the Lynn et al. (2013) subgroup comparison were analyzed alongside qualitative data from principal interviews and teacher responses on the unstructured items from the NCEES Teacher Survey. This analysis occurs in Chapter 5.

As with the previous exploration of the NCTSR movement across years, data from Weisberg et al. (2009), Curtis (2011), and Lynn et al. (2013) were standardized. This task was accomplished by taking the proportion of teachers rated across the four rating categories in the observed sources and transforming the data into equivalent sample sizes of the schools under study. A chi-square analysis then followed to detect whether there was significant statistical differences among the observed sources and the schools under study.

As with the analysis of the NCTSR movement across years, Standards 1 and 4 were used in the analysis; and Schools A, B, C, and E were involved to ensure sample size integrity within the chi-square analysis. The results of this analysis were displayed in Table 13 with the observed source listed alongside the standardized sample size, the chi-square statistic, degrees of freedom, and p values.

Table 13

2013-2014 NCTSR Chi-Square Comparison to Three Observed Sources

Observed Source	School	Standard	N	X^2	df	p
Weisberg et al. (2009)	A	1 Leadership	75	177.64271	2	0.0000***
	A	4 Practices	75	152.42307	2	0.0000***
	B	1 Leadership	63	274.25	2	0.0000***
	B	4 Practices	63	228	2	0.0000***
	C	1 Leadership	80	142.47272	2	0.0000***
	C	4 Practices	80	222.72272	2	0.0000***
	E	1 Leadership	69	230.52941	2	0.0000***
	E	4 Practices	69	239.58333	2	0.0000***
Curtis (2011)	A	1 Leadership	75	14.19364	2	0.0008***
	A	4 Practices	75	7.54818	2	0.0230*
	B	1 Leadership	63	72.69523	2	0.0000***
	B	4 Practices	63	51.38100	2	0.0000***
	C	1 Leadership	80	8.91300	2	0.0116*
	C	4 Practices	80	33.93566	2	0.0000***
	E	1 Leadership	69	42.85494	2	0.0000***
	E	4 Practices	69	43.09407	2	0.0000***
Lynn et al. (2013)	A	1 Leadership	75	15.45714	2	0.0004***
	A	4 Practices	75	24.45357	2	0.0000***
	B	1 Leadership	63	8.77189	2	0.0125*
	B	4 Practices	63	4.64764	2	0.0979
	C	1 Leadership	80	5.92652	2	0.0517
	C	4 Practices	80	4.42890	2	0.1092
	E	1 Leadership	69	1.24233	2	0.5373
	E	4 Practices	69	3.10932	2	0.2113

*Note. *denotes significant difference at* α*=0.05; **denotes significant difference at* α*=0.01; ***denotes significant difference at* α*=0.001.*

The chi-square analysis produced data showing the schools under study significantly differed from Weisberg et al.'s (2009) teacher rating distribution resulting in a rejection of the null hypothesis. The relationship between the rating distribution of the schools under study versus Weisberg et al.'s rating distribution was substantial, and all comparisons were significantly different at α=0.001 without registering a single value.

Because of the radical p values, a comparison of the chi-square test statistic (χ^2) better reflected the substantial differences between Weisberg et al.'s distribution compared to the schools under study ranging from between χ^2=142.47272 and X^2=274.25. These chi-square statistic outcomes reinforced the degree of rating inflation of historical teacher evaluation systems in Weisberg et al.'s study and also showed that NCEES has produced a rating distribution that looked signficantly different.

The rating distribution of the schools under study approached the rating distribution found in Curtis's (2011) study with Schools A and C at least registering a p value, although still significantly different. An analysis of the chi-square statistic (χ^2) better reflected the decrease in the significance of the differences between the schools under study compared to that of Weisberg et al. (2009) and Curtis. When the schools under study were compared to Curtis's outcomes, the χ^2 ranged from between χ^2= 7.54818 and 72.69523, substantially lower than the χ^2 values from the comparison with Weisberg et al.

The rating distribution of NCEES for the schools under study were better aligned with the rating distribution of Lynn et al. (2013) with most registering nonsignificant differences. The χ^2 values also reflected the more approximate relationship between the NCEES teacher ratings for the schools under study and Lynn et al. compared to that of Weisberg et al. (2009) and Curtis (2011).

After further investigation, it was found that the NCEES teacher ratings for the schools under study were more tightly clustered than Lynn et al. (2013) data. Observers in the schools under study rated teachers "distinguished" and "developed" at about half the rate as the observers in Lynn et al., which the authors deemed "clustered" (p. 24) around the two middle ratings.

The ratings of the schools under study appear to be clustered to a higher extent especially within the "accomplished" category. Rather than clustered, a more accurate description may be that the ratings of the schools under study have been *compressed* into the "proficient" and "accomplished" ratings, especially the latter accounting for 57% of ratings. The "compression" description is evident in Figure 9, which graphically shows the overall proportional differences among the teacher ratings in the schools under study and the three observed sources.

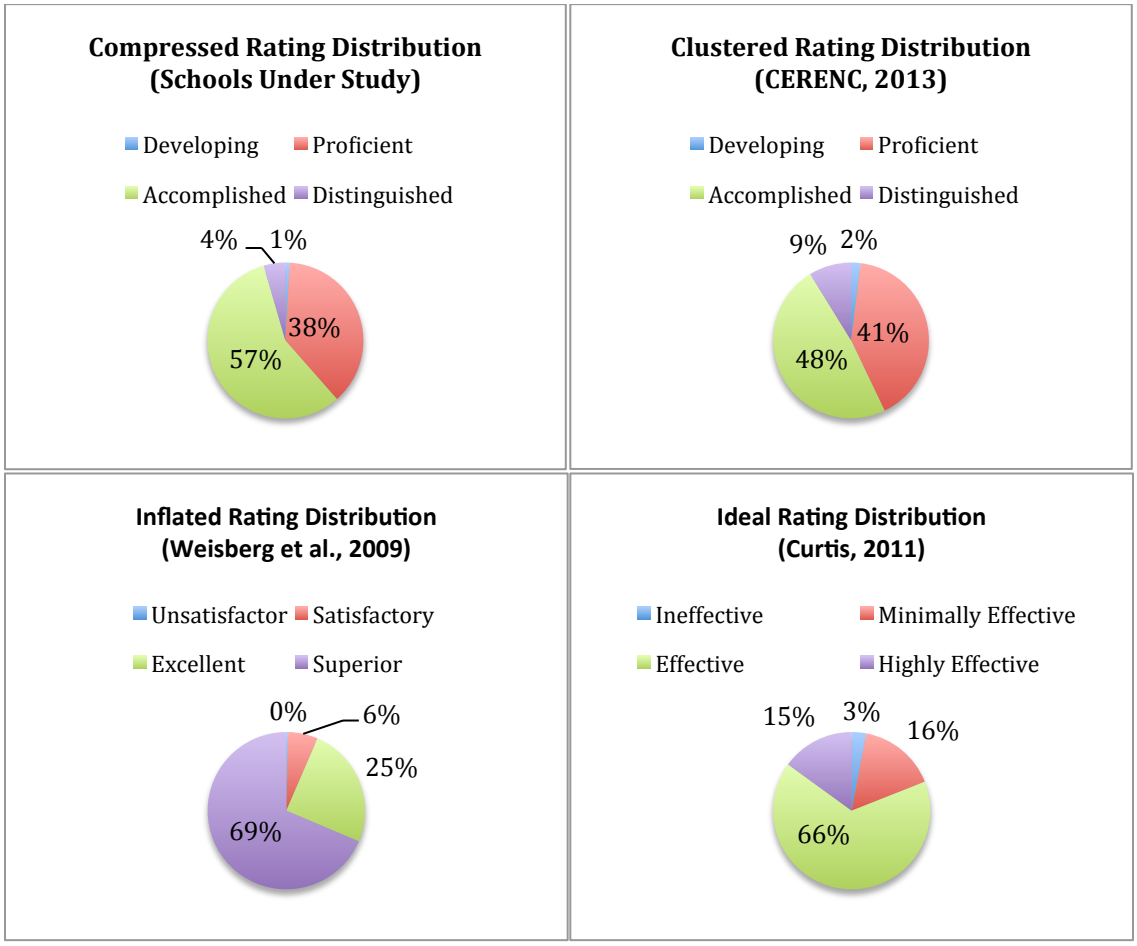


Figure 9. Rating Distributions for Schools under Study and Three Observed Sources.

As with the other measures involving the NCTSR, it was important to triangulate the quantitative outcomes with the qualitative input from principals and teachers to

measure the effects of the compression of ratings on teaching practices and teacher leadership. This approach is taken in Chapter 5.

In summary, the quantitative methodology of this study has shown that teachers responded positively to 13 items on NCEES versus five in which there was negativity. The negative responses from teachers centered on the validity and reliability of EVAAS® outcomes, the quality of formative feedback received from observers, and the ability of summative ratings to improve teaching practices. An analysis of the measures of central tendency reinforced a previous analysis of positive-to-negative and negative-to-positive ratios on the NCEES Teacher Survey.

Statistical tests showed there to be significant differences among certain groups of teachers on items from the NCEES Teacher Survey. Teachers with less experience seemed to agree at a higher rate of proportionality that they had adequate time to participate in pre and postconferences; teachers with less experience seemed to agree at a higher rate of proportionality that various observers showed agreement in rating their performance for Standards 1-5; teachers with more experience and who had experienced a previous teacher evaluation system seemed to agree at a higher rate proportionally that Standard 6 ratings reflected the academic growth of their students; teachers with less experience and who had not experienced a previous teacher evaluation system seemed to agree at a higher rate proportionally that they received useful feedback from observers to improve their performance.

Statistical tests were used to describe the NCTSR across years and across three observed sources. Using the statistical tests, it was shown that the NCTSR had significantly changed between the 2011-2012 and 2013-2014 school years for Standards 1-5. Also, Standard 6 also showed statistically significant movement between the 2012-

2013 and 2013-2014 school years for two of the four schools under study. After further investigation, it was determined that the NCTSR for the schools under study had migrated towards the "proficient" rating category.

NCTSR was also analyzed against three observed sources chosen for their status in the literature. NCTSR for the schools under study were shown to be significantly different than those found in Weisberg et al. (2009) and Curtis (2011). Also, it was shown that NCTSR for the schools under study were clustered around the two middle ratings of "proficient" and "accomplished" to a higher degree than what was found in Lynn et al. (2013) leading to a compression of ratings in the middle.

Qualititative Findings

The qualitative data that were collected included input from teachers on the unstructured items on the NCEES Teacher Survey and from principals on the NCEES Principal Interview. The teacher data were collected for each item on the NCEES Teacher Survey and included all the demographic details of the respondents. The demographic details included the school in which the teacher taught, the teacher's gender and ethnicity/race, the teacher's years of teaching experience, and whether the teacher had experienced a previous teacher evaluation system. The demographics of the principals were also collected at the time of each interview and included their years of experience as a teacher and as an administrator, alongside information about whether they had experienced a previous teacher evaluation system as a teacher or an administrator.

The principals included a wide variety of individuals who experienced NCEES at many different levels. Two principals had more historical experience with teacher evaluation systems in North Carolina, with one specifically dating back to the 1990s and the other as a teacher receiving classroom observations and evaluation using the former

North Carolina instrument (TPAI-R) and then using NCEES as a principal to observe and evaluate teachers. The remaining four principals experienced receiving classroom observations and evaluation using NCEES as teachers and then using NCEES in an administrative capacity in observing and evaluating teachers in their respective schools.

For the qualitative analysis, grounded theory was used following Charmaz's (2006) coding, generation of theoretical concepts from testing connectedness among categories and memos, and theory building. Initial coding involved highlighting each word, phrase, or sentence for relevance and importance; and a focused coding process followed that attempted to sort the data into pieces of evidence that summarized and possibly supported an idea anchoring the concepts and processes involved. The codes were then elevated to overarching memos based on relevance and importance. The memos were analyzed for interconnectedness and elevated to categories used to build concepts and theories about how NCEES has affected teaching practices and teacher leadership in the schools under study.

It was important for the qualitative data to act as a stand-alone source of data apart from the quantitative data to objectively consider the perceptions of teachers and administrators in describing the effects of NCEES on teaching practices and teacher leadership. However, another important goal of analyzing the qualitative data was to help understand the effects of the quantitative data from the NCEES Teacher Survey, NCTSR movement across years, and the NCTSR comparison across the three observed sources. This approach would eventually lead to triangulation.

Synthesis of Novel Categories and Conceptual Questions Using Grounded Theory

An initial and focused coding process took place without the intention of generating novel categories, connected conceptual questions, or theories compared to

what was discovered in the pilot data (collected 2 years previous) and the literature review. However, there was one novel category and several novel connected conceptual questions within existing cateogries that were built using qualitiative data from the primary research compared to the secondary research from pilot data and literature review.

There were similar reoccuring codes that eventually diverged from the categories presented in Figure 2 from the pilot data and literature review. Similar codes combined at the lowest level of initial and focused coding, transferred to memos, and ultimately formed categories or extended conceptual questons within existing categories that could be investigated in future studies. Although an attempt was made during this analysis to build theories pertaining to the concepts and processes of this newly devised category and extending conceptual questions, data were limited because the NCEES Teacher Survey and NCEES Principal Interview were not designed to measure this novel category or novel conceptual questions. Future studies could center on building theories to explain how this novel category and extending conceptual questions work within teacher evaluation systems and their effects on teaching practices and teacher leadership.

The first novel category related to the outcomes of NCEES but differed from the context and meaning of the already existing category in Table 2, “How Are Outcomes From Teacher Evaluation Systems Utilized?” The memo title, “Teacher and Principal Expectations of Outcomes,” was chosen to represent the many pieces of evidence collected in the initial and focused coding procedures.

The novelty that set this new category and surrounding concepts and processes apart was that it carried a different contextual status than how the outcomes from NCEES were utilized. The category, “Teacher and Principal Expectations of Outcomes,” reflected

mental processes, whereas the previous category pertained to the structure and mechanics of what occured during and after preconferences and postconferences and how outcomes from teacher evaluation systems were being used to systematically improve teaching practices and teacher leadership.

The choice to present a new category into Figure 2 was based on the differences in meaning where the coding eventually led. To eliminate the novel category and break down its meaning into additional conceptual questions that could be placed in the already existing category, “How Are Outcomes from Teacher Evaluation Systems Utilized,” would not have been justified given the different meaning between the top, axial coding levels of memos and the category. There was a meaningful difference about the memos and resulting category, “Teacher and Principal Expectations of Outcomes,” at the memo and categorical level that demanded it be contrasted with the existing category, “How Are Outcomes from Teacher Evaluation Systems Utilized?”

Evidence from the codes collected for this novel category presented concepts that placed an important philosophical distinction between principals and teachers before attempting to determine how the outcomes from NCEES were being used to improve teaching practices and teacher leadership. The codes pointed to the kind of feedback teachers were expecting to receive to improve teaching practices versus what principals aimed to provide. The novel category was conceputally built on preexisting mindsets of teachers and principals before ever entering into communciation in pre and postconferences. The codes pointed to a philosophical difference between what feedback principals believed was their duty to provide and the feedback teachers expected to receive, and ultimately this philophical difference acted as a barrier to inititating improvement in teaching practices.

Evidence existed in the coding of teacher responses for Item 35 on the unstructured portion of NCEES that indicated teachers expected specific feedback from principals as to how they should improve teaching practices. During interviews, principals uniformly expressed a desire to cultivate a reflective state among teachers to preserve teacher professionalism and spur teachers to think for themselves about how they could improve teaching practices. The responses of teachers and principals displayed the contrasting philosophies apparent between them that drove a wedge between principals and teachers and how they interacted during times of feedback. The evidence provided by teachers in their responses caused teachers to view feedback from principals in the following ways:

- There was not enough feedback.
- The feedback was not as specific as teachers desired.
- The feedback did not match expectations teachers held for their particular subject area.
- The feedback that teachers received during the summative evaluation was not considered applicable because it was not from a principal that observed their teaching practices throughout the school year.

There were a few comments from teachers who received little or no feedback from observers, but these responses were rare. The majority of teacher responses revealed they received feedback; but the feedback was not adquate because it was not enough, not specific, or it did not apply to their subject area. Such comments included,

- I have not received feedback other than what boxes were checked on the observation. For an observation to be useful to the teacher, the teacher needs to

receive specific feedback on the lesson not just checks on a form.

- I have received compliments about things that I do well, but never really any insight into relevant things that I could do better to improve learning. Most of the time, the administrator who is observing me is not knowledgeable in my subject, has not taught long enough to provide useful advice, or has no time to spare really anlalyzing my teaching ability or practices. Basically observer feedback rarely affects my teaching practices.
- Although I am receiving more feedback than ever, it is generic. Principals don't know my subject because it is a specialty area, and they don't understand the struggles I encounter. I rely more on those around me.
- I appreciate the feedback received during these conferences but get frustrated when the feedback is limited or generic.
- As a new teacher . . . I need to hear more about what to do in specific situations. Sometimes I get feedback that isn't what I need even though I'm told about places I need to improve. Feedback is not specific to what I'm told I need to do better.
- The administrator that did my summative had not even been in my room to observe me this year. How can he give me feedback?
- Give me places to look at how to improve or to help me tweak my approach.
- Principals do not give specific feedback because they don't really get what my students are doing to learn. I can walk through a power plant many times but still have no idea how electricity is made.

The codes that sythesized the previous statements fit into the outcomes category

in Table 2; but when contrasted with statements with principals, there was a cause-and-effect relationship that demanded a novel category in which theories could extend in this study and possibly in the future. While teachers expected more and better feedback regarding teaching practices, principals expected teachers to use their inherent skills and knowledge of teaching to generate ways to become better. Rather than walking "through a power plant many times" attempting to understand how electricity is made and directing power plant workers in ways to do it better, principals have been attempting to use and share the results of low-inference observations to ensure teachers have an "idea" of "how electricity is made" and act upon that knowledge to produce electricity (i.e., learning) in a more productive way. The evidence provided by principals in their responses revealed a philosophical difference in their approach to providing feedback that could be summarized by

- a relunctance to infringe on classroom pedagogy;
- a desire to develop a teacher's sense of problem-solving skills to improve teaching practices;
- using peers to help improve teaching practices; and
- protecting and attempting to cultivate teacher professionalism to improve teaching practices.

In interviews, principals made the following comments revealing their desire to act as a facilitator driving teachers to reflect on their practices with the intent of improving teaching practices and teacher leadership:

- One of the things we concentrate on is getting teachers to think about how they could do things better. I'll tell them where they need improvement. I'll

tell them about things I can see . . . I'll be open. But teachers know what they do day in and day out much better than anyone and they can be a big source of improvement . . . and I can help them get there.

- Reflection is Standard 5. If the conversation is right, teachers here can reflect and identify how to improve the way they engage students. Engagement is important here because of our at-risk kids. We work together to find things that work. I've worked with this population long enough to know engagement is key and if we can't figure out some things together to get that engagement piece . . . I can send them to observe other clasrooms to teachers who do this on a regular basis. I have done that, and it works.
- With the old instrument everybody was used to being at or above standard. This new instrument has surprised some veteran teachers . . . I coach them through it. Sometimes its tense. But they've got what it takes . . . they want to do well and be accomplished and distinguished, or they probably wouldn't be here to start with. I want to help them get there, and even if there is disagreement, I want to hear their side and see artifacts. I always stay open to this type of conversation because it causes them to think about how they could do things better and be professionals that can get better and come up with solutions.
- Most of the time they are listening and thinking about how to get better.
- When students are learning . . . and now we can see it with different data . . . with their growth index and things we are observing, we still want improvement. And I have told teachers in the past that their growth scores are

good and we are proud of the work they do. And if you look, these teachers are the ones thinking about things they do, and they are innovative. They have solid PDP's, and they are constantly looking for ways to get better and make others around them better.

- These were hand-picked teachers. It was a small crew. They are really strong so I did not have to tell them how to do things because they already knew and were doing them. Our growth scores were indicators. For me it was a matter of respecting their professional training and allowing them to do what they do best.

Principals in the county under study provided feedback to teachers in a manner parallel to Danielson's (2010/2011) approach–to promote "self-assessment, reflection on practice, and professional conversation" (p. 38). Principal responses to the NCEES Principal Interview process generated codes as synthesized by a grounded theory analysis that were very similar in content to Danielson's (2010/2011) approach. Danielson (2010/2011) described the traditional methods of observing and evaluating teachers as inadequate because the teacher was not an active participant in the observation and evaluation process.

Danielson (2010/2011) provided contrasting modes of providing feedback within traditional teacher evaluation systems compared to revamped teacher evaluation systems. The traditional approach posited the teachers as a passive participant:

> We can get a clue as to the nature of this problem if we consider the typical observation, supervision, and evaluation process in use in most schools. The scenario proceeds as follows: The administrator goes into the classroom and watches a lesson, takes notes, goes away and writes up the notes, and then returns

> and tells the teacher about the lesson (what was good, what the teacher could improve). Most observations are a variation of this theme.
>
> It's important to note that in this scenario, the administrator is doing all the work; the teacher is complete passive. (The teacher has, of course, taught the lesson, but the teacher contributes nothing to the observation itself). So it's not surprising that teachers don't find the process valuable or supportive of their learning. The process violates everything we know about learning–that learning is done by the *learner* through a process of active intellectual engagement. (Danielson, 2010/2011, pp. 37-38)

Danielson (2010/2011) contrasted the traditional approach with an approach found in revamped teacher evaluation systems, including NCEES as indicated by principals in their responses within the NCEES Principal Interview process:

1. The administrator goes to the classroom, watches a lesson, and takes notes on all aspects of the lesson: what the teacher says and does, what the students say and do, the appearance of the classroom, and so on.
2. The administrator gives a copy of his or her notes to the teacher.
3. The administrator analyzes the notes against the evaluative criteria and levels of performance.
4. The teacher reflects on the lesson using the observer's notes and assesses the lesson against the evaluative criteria and levels of performance. The teacher will probably, as a result of this reflection, identify aspects of his or her teaching to strengthen, and that teacher will reach these conclusions without prompting from the principal. Of course, the principal can always point things out, but when the teacher reflects on a lesson before the postobservation

conference, he or she will frequently be as critical as the principal would have been.

5. The teacher and administrator discuss the lesson. The teacher puts the lesson into context for the administrator, and together they decide on the teacher's strengths and areas for growth. Naturally, the administrator wasn't in the classroom the previous day and can't be familiar with all the issues that the teacher must address. So the teacher might describe a particular student's learning challenges, and the principal might suggest a different approach. But they conduct the conversation in light of their shared understanding of what constitutes good teaching. (p. 38)

Blase and Blase (2000) surveyed 809 teachers and examimed how teachers viewed principals' instructional leadership. The teachers responded in similar fashion to Danielson's (2010/2011) approach. Blase and Blase summarized the approach of successful principals in improving teaching practices and ultimatlely student achievement:

> We found that in effective principal-teacher interaction about instruction, processes such as inquiry, reflection, exploration, and experimentation result; teachers build repertoires of flexible alternatives rather than collecting rigid teaching procedures and methods. Our model of effective instructional leadership was derived directly from the data; it consists of the two major themes: talking with teachers to promote reflection and promoting professional growth. (p. 3)

While principals in the county under study generally followed Blase and Blase's (2000) and Danielson's (2010/2011) approaches to observing and evaluating teacher performance and offering useable feedback, teachers indicated they desired more specific

feedback. Future studies would be necessary to determine where the differences were to be found between the threshold of specificity to which principals offered feedback and the level of specificity to which teachers desired.

To summarize, teachers expressed a desire to receive specific feedback based on their performance, but the approach of principals was to offer generalized feedback on the performance of teachers and rely on teachers to analyze their teaching practices for areas of improvement and to act upon their analysis. Principals wanted to be agents of change within the process as facilitators of improvement without overpowering the professional status of teachers and respecting their ability to solve problems. New conceptual questions, "What are the goals of principals in offering feedback to teachers about their teaching practices and teacher leadership" and "What are the goals of teachers in receiving feedback from principals about their teaching practices and teacher leadership," were placed in a new category, "Principal and Teacher Expectations of Feedback."

It would seem that with two sets of qualitative data collected 2 years apart (pilot data were collected 2 years earlier), there would have been more discrepancies among codes, memos, categories, and derived theories between the two sets of data. There were two possible explanations for why there were few discrepancies.

First, it was possible given the scope of the literature review and its influence on designing the data collection instruments that the categories approached saturation with concepts that identified the majority of issues surrounding teacher evaluation systems and their effects on teaching practices and teacher leadership. The second reason was connected to the actual design of the instruments used to collect data.

The responses from the teachers were based upon specific items on the NCEES

Teacher Survey designed from the pilot data and literature review, and there was little deviation in their answers in terms of migrating away from the original intent the item on the NCEES Teacher Survey was measuring. Although considered "unstructured" because of the ability of teachers to provide written responses rather than respond on a Likert scale, the wording in the items on the NCEES Teacher Survey served as goalposts limiting the breadth of responses among teachers. As a result, there was only one set of codes that translated into higher-level memos and eventually the category "Teacher Expectations of Outcomes," which was then combined into "Teacher and Principal Expectations of Outcomes" after considering the difference in expectations of outcomes from principals.

For the principals, their responses within interviews were more open, but the NCEES Principal Interview questions still served as minor goalposts. Principals often migrated away from the intent of the original interview question revealing important perceptions about processes involving NCEES which were indirectly related to the original interview question.

From within the responses from principals, there were two novel conceptual questions placed within the existing category, "Quality of Classroom Observation and Teacher Evaluation Within Teacher Evaluation Systems," as depicted in Figure 2. The novel conceptual questions did not carry with them different meanings at the memo level that demanded they form a new category. These novel conceptual questions created concepts that could be further investigated to form new theories about how teacher evaluation systems affect teaching practices and teacher leadership. The novel conceptual questions included

1. the expectations of outcomes, which were already examined and placed into a

category named “Principal and Teacher Expectations of Feedback”;

2. the level of training that principals received and how the depth of training affected their ability to carry out the processes involved in observing and evaluating teachers in using NCEES; and
3. the support mechanisms available to principals that provided aid in using NCEES.

The codes that involved the training of principals reflected concepts and theories within the previously designed category entitled “Quality of Classroom Observation and Teacher Evaluation Within Teacher Evaluation Systems,” found within Figure 2 in Chapter 2. Within this category, there was specific attention in the literature review given to the importance of calibrating raters in teacher evaluation systems for the sake of reliability. Upon further inspection, the responses of the principals provided deeper concepts and processes than calibration efforts encompassed. As a result, there needed to be new questions sythesized that addressed the training of principals and the support mechanisms of which they made use.

Principals would frequently respond that because of their training, they chose a certain course or handled situations involving NCEES in a predesigned fashion that stemmed from their training. Principals were specific in their responses indicating that it was because of being trained and the professional development they received on NCEES that they

- knew the advantages of informal visits to classrooms to improve teaching practices and teacher leadership;
- believed it was as important to provide feedback to veteran teachers for

improvement as it was to provide feedback for improvement to new teachers;

- attempted to control rating inflation and knew how to avoid it;
- viewed NCEES as a tool that has special status because of the time, money, and energy it required to institute;
- believed NCEES was a growth instrument used to garner improvement they wanted to see in teachers over time;
- thought it was important that multiple observers be used to observe and evaluate teachers in order to improve teaching practices and teacher leadership;
- viewed the "Proficient" rating as teachers performing at a satisfactory level, while "Accomplished" and "Distinguished" ratings were reserved for teachers who geared their instruction and leadership towards higher levels of collaboration among stakeholders;
- accepted the validity of EVAAS® outcomes as their understanding of the model increased; and
- were concerned about a portion of their training mandating they only rate teachers based on what they see during a classroom observation and the effects of that mandate on teaching practices and teacher leadership.

While all the phrases and wording from these statements formed codes that fit congruently into previously developed categories and theories from the pilot data and literature review, it was the straightforward admission that these beliefs were an extension of the training principals received that set these codes apart. After mentioning that their specific actions resulted from training, principals would naturally expound on

the positive effects of their actions on teaching practices and teacher leadership.

Principals recollected and oftentimes reminisced that their training had influenced their ability to operate while using NCEES. The following comments served as the embodiment of the training principals received, the actions they took as a result, and the effects on teaching practices and teacher leadership:

- I went to a training . . . we were shown the advantages of informal visits . . . in visiting classes time after time, and it helped teachers build a rapport [with me] and allowed me to give them a sense of direction . . . they would accept.
- We have been told by trainers that this system is designed for everyone, not just beginning teachers or those that would obviously need help . . . struggling teachers . . . so we offer feedback to veteran teachers, and I think they appreciate that most of the time. Most of the time they are listening and thinking about how to get better.
- We split up the observation schedule . . . we were told to do this in training . . . we get more than one administrator in there. This helps get a uniform message out there about what it takes to get better.
- The amount . . . and the quality . . . of training with this system was better. I can remember in the 90s, and I can remember ten years ago . . . the training was useful. And I think we used it, and I'm not saying principals back then didn't take stock. But this new system . . . you could tell with the training, this is for real. It has teeth to it, something we didn't have ten and twenty years ago. The time and money . . . the manpower it took . . . you can really tell with the training sessions I've been to they want it to matter.

- It used to be about the snapshot . . . the slice of time . . . now its more about, "What are you showing over time?" "Where were you when you started this?" "Let's look at your PDP and talk about what you're doing to improve the level of instruction and engagement for students." This is what we're told at training to communicate to our teachers so they know, and I think talking about this information has helped.
- We have been told many times in training that when teachers are doing what's expected and at levels where kids are learning . . . the environment is stable . . . and everything's going right, that there's nothing wrong with being proficient. When they're [teachers] aware and they know and do it, they are proficient. Teachers are evolving . . . I think they get now that it's going to take more to be accomplished or distinguished. Most [teachers] are set with that plan, and some go that extra mile to help their colleagues help their own students.
- Every time I go to an EVAAS® training I learn more . . . which I'm set to go to another soon. Every time I go I feel more comfortable . . . and this goes with talking to teachers. I think there are still teachers out there that see it as a bad thing because of teaching certain kids . . . but because of the training I've gone to I can explain that this is about growth not their [students'] overall score. I think teachers are getting that.
- We're told in training to mark only what you see in the classroom . . . so it doesn't accurately portray what's going on in that day. I see a problem here . . . a contradiction . . . so now there's inflation because I know this teacher is

doing this stuff, just not right now . . . and the credibility can come into question.

The second novel coding outcome that eventually filtered up into conceptual levels within the existing category, "Quality of Classroom Observation and Teacher Evaluation Within Teacher Evaluation Systems," was the extension of support mechanisms available to principals in their effort to use NCEES effectively to improve teaching practices and teacher leadership. Principals communicated that they relied on county personnel, online resources, and each other in order to refine their understanding of the NCEES standards and processes. Principals resonded that it was because of support mechanisms that they

- were empowered to make suggestions to teachers who would improve teaching practices and teacher leadership;
- could receive solution-based responses from county personnel to avoid potential problems;
- tapped into the knowledge of their peers in order to solve misconceptions both at the teacher and principal level which resulted in improving teaching practices and teacher leadership; and
- rated teachers with increased reliability and provided feedback to teachers who improved their teaching practices and teacher leadership.

The following comments reflected the reliance of principals on support mechanisms to more effectively use NCEES, the results of receiving support, and the effects on teaching practices and teacher leadership:

- Some times there's been confusion over what something means on . . . like

Standard 2. So we've gotten together with a teacher and other teachers and looked at things they could do to implement diversity in a lesson. This has really helped improve instruction to get people talking about what to do and how to do it.

- I knew that the county office was always a phone call away at that time and– really helped. The rating ["not demonstrated"] had confused some people . . . we were able to get it ironed out.
- The county people have been just steps away. And there were times where when I needed them, they walked over here and observed teachers. I really used that several times . . . they came to me . . . and I plan to continue it.
- I have found answers online within the website. At first I think we all had problems navigating . . . and some things changed with certain radio buttons and functions. But the access to data is so much better than with any system I used before.
- Its been important for me as a newer administrator to be able to talk to [other administrators] . . . the discussions we had at the beginning really helped. I still had to go back and look and talk to [other administrators] to make sure we were seeing things the same way. And we talk a good bit about things we've seen and the level of feedback we need to give to see improvement. And that's important.

Like the coding for training principals received, the resulting codes and concepts surrounding support mechanisms fit neatly within the previously developed category, "Quality of Classroom Observation and Teacher Evaluation Within Teacher Evaluation

Systems" in Table 2. The concepts surrounding the use of support mechanisms warranted new conceptual questions to be intertwined in future theory building. A new version of the flow between existing and the novel category alongside novel conceptual questions was redesigned as displayed in Figure 10 adapted from Figure 2.

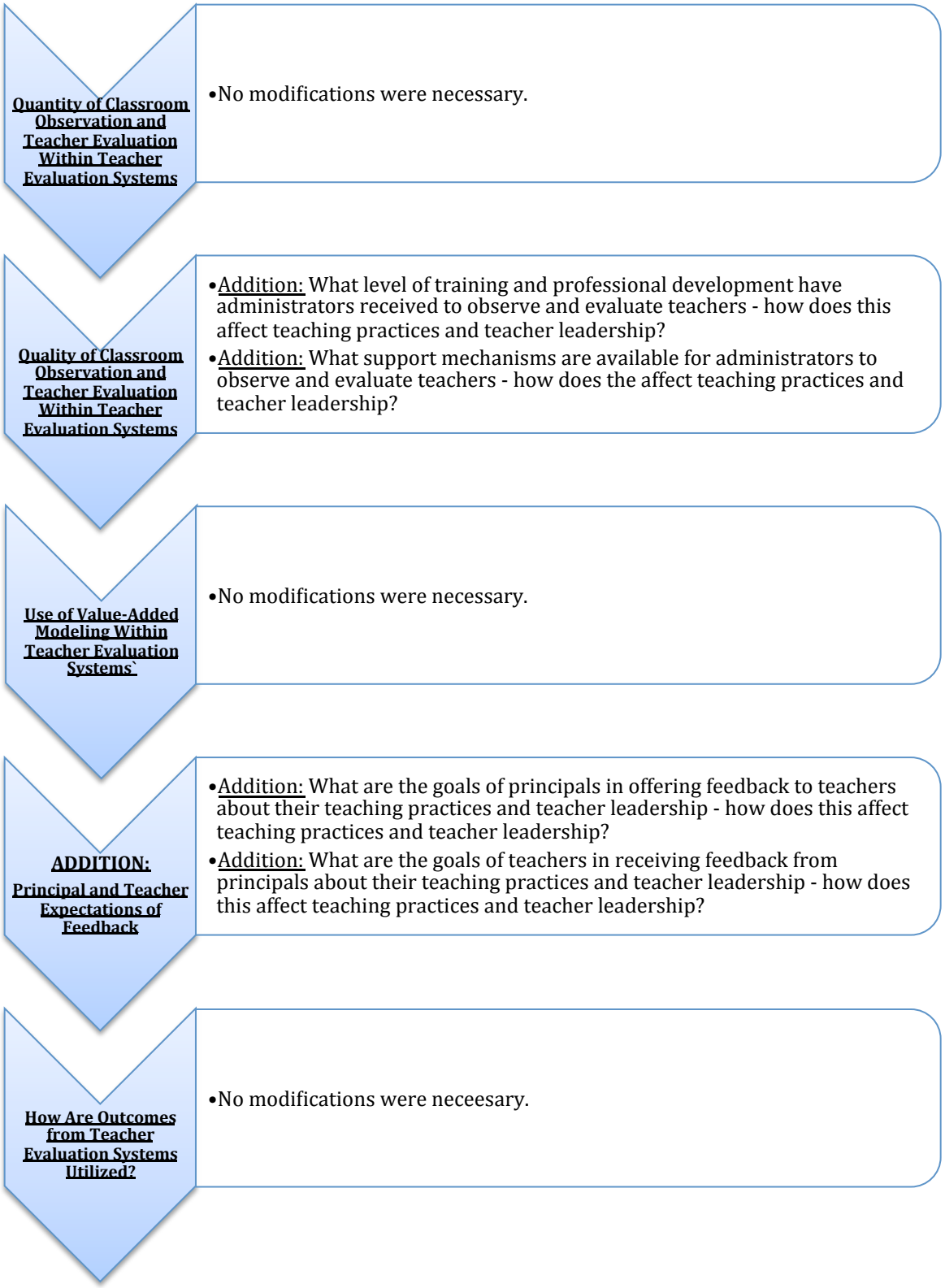


Figure 10. Adapted Figure 2–Novel Category and Subcategory Concept Questions.

Analysis of Unstructured Responses in Existing Categories Using Grounded Theory

Quantity of classroom observation and teacher evaluation within NCEES. In measuring structured responses on the NCEES Teacher Survey for the category "Quantity of Classroom Observations and Teacher Evaluation Within Teacher Evaluation Systems," Items 7, 9, and 11 were designed to measure teacher perceptions of whether they received adequate formal and informal classroom observations and whether there was adequate time availalble to participate in pre and postconferences. Items 7, 9, and 11 were overwhelmingly positive with positive-to-negative ratios of 3.63, 2.07, and 2.40, respectively (see Table 8).

Unstructured Items 8 and 10 gave teachers the opporunity to share their perceptions about whether they received adequate formal and informal classroom observations and how their teaching practices and teacher leadership were affected as a result. Unstructured Item 12 gave teachers the opportunity to share their perceptions of whether there was an adequate amount of time to participate in pre and postconferences and how their teaching practices and teacher leadership were affected as a result (see Appendix A).

Teachers appreciated the effects that consistent formal and informal classroom observations had on their teaching practices and teacher leadership in unstructured Items 8 and 10. Teachers communicated that informal walkthroughs "kept them on their toes" making sure their "I can . . . statements" were posted and their organized lesson plans were on hand. Both unstructured Items 8 and 10 contained teacher free responses signifying that consistent formal and informal classroom observations caused teachers to plan their lessons with student engagement and learning as high priorities. There was also a sense from teachers that they appreciated informal observations because it kept both

themselves and their peers accountable for providing quality teaching practices and teacher leadership.

When analyzing structured responses across Items 7 and 9 (see Table 8), there was a noticeable decrease in the number of teachers who agreed or strongly agreed they had received an adequate number of informal classroom observations (positive-to-negative ratio=2.07) compared to teachers who agreed or strongly agreed that they had received an adequate number of formal classroom observations (positive-to-negative ratio=3.63). There were some responses on unstructured Items 8 and 12 that when coded provided evidence for changes in this perception.

An investigation ensued centering on teachers who agreed or strongly agreed in Item 7 (adequate number of formal observations) and then in turn disagreed or strongly disagreed in Item 9 (adequate number of informal observations). The investigtion was carried out by analyzing unstructured responses for Items 8 and 10 in an attempt to discover why the change in perception had occurred across Items 7 and 9.

Teachers who changed in their perception from “agree” and “strongly agree” downgraded their ratings to “disagree” and “strongly disagree” across Items 7 and 9 because they questioned the ability to collect worthwhile amounts of data in such a short amount of time experienced in an informal classroom observation. As a result, teachers perceived that they received less useable feedback, and these teachers challenged the number of informal classroom observations they received. This was a similar sentiment shared by teachers in unstructured Items 8 and 10 who disagreed or strongly disagreed across both structured Items 7 and 9.

Some teachers who disagreed and strongly disagreed in structured Items 7 and 9 stated in unstructured Items 8 and 10 that they had not received an adequate number of

formal observations or informal observations in terms such as, “How can one formal observation fully describe what I do” and “I only received one walkthrough at the beginning of the year.” In most cases, this response was not followed with a full description of the effects on their teaching practices and teacher leadership; but there was an overall suspicion that without enough observation data, they would not receive quality feedback.

There were a small number of responses on unstructured Item 10 that communicated these teachers had not received walkthroughs or snapshots at all; although there were no teachers who responded on unstructured Item 8 that they had not received the prescribed number of classroom observations based on their status as a beginning, career teacher, or renewing career teacher. This was an important revelation because it was mandated under NCEES that all teachers receive a certain number of classroom observations based on their aforemetioned status (McREL et al., 2012), and teachers agreed they were receiving them. This noted an improvement in NCEES where principals collected ample data about teacher performance compared to traditional teacher evaluation systems where teachers were rarely observed and evaluated.

Comments on the unstructured portion for Item 10 ranged from “with the informals they may only see me lecture and that is NOT what my class is about” to “I don’t think any of my snapshots or walkthroughs have been recorded in NCEES system”; and “Because I don’t receive much feedback from those [informal observations], they do not significantly affect my teaching practices or leadership.” It appeared from the responses on the unstructured portions for Item 10 that teachers who disagreed or strongly disagreed held contrary perceptions about the use of informal classroom observations compared to evidence provided by principals in their interviews.

Principals uniformally communicated in their interviews that informal observations provided them the ability to collect data about the general patterns in teaching practices and leadership activities that were occurring in classrooms. One principal replied that informal classroom observations were used in large part to check that teachers had a focused lesson plan with "I can . . . " statements posted for students. Principals also responded that checking student engagement and the use of technology encompassed a large motivation for carrying out walkthroughs and snapshots.

In summary, although a handful of teachers contended they either did not receive snapshots or walkthroughs or received a minimal number and some teachers communicated a mistrust in informal observations because of the small amount of data collected, the vast majority of teachers agreed or strongly agreed that they had received an adequate number of formal and informal classroom observations. The effects of adequate formal and informal classroom observations on teaching practices and teacher leadership included improved lesson planning, keeping students focused and engaged, and teachers providing their students quality teaching practices and teacher leadership because of a level of accountability.

The positive-to-negative ratio for structured Item 11 was 2.40 (see Table 8), and teachers were in agreement overall that there was enough time to participate in pre and postconferences. For structured Item 11, teachers agreeing or strongly agreeing they had an adequate amount of time to participate in pre and postconferences relayed the positive effects on their teaching practices and teacher leadership. The main finding in the codes was that with adequate time for principals and teachers, teachers received adequate feedback about what they were doing well, where they needed to improve, and that they had time to reflect and plan to improve their teaching practices and teacher leadership.

There was some evidence found in coding Item 12 there was a shortage of time for both teachers and principals, including perceptions from teachers who principals did not have the time necessary to provide quality feedback. This perception was not overwhelming and did not appear to carry as much consistency or importance as in the pilot data, especially among principals. In the pilot data, principals centered much of their discussion on the lack of time to carry out the observation and evaluation process with fidelity. The literature review also provided examples of teachers and principals struggling to operate within recently revamped teacher evaluation systems because of a shortage of time. This perception was not as apparent in the primary research.

When teachers disagreed or strongly disagreed for structured Item 11, some teachers followed with descriptions in unstructured Item 12 about how an inadequate amount of time they experienced negatively affected their teaching practices and teacher leadership. There was evidence provided in unstructured Item 12 that teachers did not have enough time to compile artifacts to demonstrate an "accomplished" or "distinguished" rating, and the teachers would settle for "proficient" ratings. Teachers communicated that it was not worth the time to collect artifacts.

Some teachers also expressed being rushed through postconferences because of a lack of time experienced by themselves and their principals. As a result of being rushed through postconferences, teachers responded that their teaching practices and teacher leadership were not improved. Teachers mentioned the following:

- More could be done with more time. I collected some artifacts to prove I'm accomplished and it took too much time.
- I can't recall ever having a preconference before an observation.

Postconferences are short and sweet because of time with very little information that could help my teaching practices.

- It hasn't changed my classroom. I feel that if the administration trusts you're doing your job, they hurry through these things to save on time.

Principals reinforced some of these practices by stating they were aware that some teachers were willing to accept a "proficient" rating because teachers did not have time to compile artifacts. Principals expressed a desire to see these teachers improve to higher ratings but also expressed hesitancy in causing an imbalance in their relationship with teachers. Two different principals commented with the following,

- When I meet with teachers the number one goal is improvement. Some teachers are going to be stressed about being "proficient." These teachers will go and find ways they are better ["accomplished" or "distinguished"]. I always stay open to these kinds of conversations because I know teachers are pushed on time. Even if they don't find ways or anything showing me they are better, I still aim to talk and figure out ways to make them better.
- Some teachers come through here and just want to sign off. They are happy with "proficient." It comes down to their leadership, and this is something we'd like to see improve without any standing bitterness. These things can be worked out with the right kind of conversations, and I wonder sometimes if they are better than "proficient" and [I] second guess my rating. But its on the shoulders of our teachers to step up with evidence to sway us which can really cause improvement in the end anyways.

One theory within this category was generated about why teachers and principals

responded distinctly different than in the pilot data or literature review in terms of the adequacy of formal and informal classroom observations and the time necessary to carry them out. Principals purposely used assistant principals and support personnel to delegate responsibilities associated with formal and informal classroom observations. Some of these supports were available only during the 2014-2015 school year and were not noticeably available to the principals in the pilot data or literature review.

For instance, one school had an extra assistant principal serving an internship, and the principal purposely used this assistant principal to help carry out consistent classroom observations. Another school had an extra assistant principal position compared to other schools with larger student populations. This extra assistant principal carried out extensive classroom observations and provided abundant feedback to teachers.

At another school, the principal made use of a ninth-grade coordinator who normally would not have been utilized to carry out classroom observations and provide evaluation input. The principal communicated that he chose to use this individual to carry out classroom observations and provide evaluation because he held credentials and had received extensive training to do so. According to this principal, his previous ninth-grade acadmy coordinators did not hold such a certification and were not used in such a capacity.

Finally, a principal at another school had a high school campus on site at the county office, and this principal made use of district personnel to provide classroom observation and teacher evaluation when necessary. District personnel were within walking distance and provided classroom observations and offered data to the principal in times of need.

Taken together, these extra support positions may have provided principals in the

county under study the ability to maximize the consistency of formal and informal classroom observations. For teachers, this may have resulted in an increase in the consistency of formal and informal classroom observations. Also, the ability to use these supports may have eased the time requirements for principals.

Teachers and principals responded with different perceptions for this category within the primary study compared to the secondary. Teachers and principals overall expressed that there were an adequate number of formal and informal classroom observations occurring, with few exceptions present. One principal accepted that NCEES required more time than the previous teacher evaluation system within which he had worked previously, but the principal did not communciate that the time necessary was overbearing or crowding his ability to accomplish other responsibilities. This principal also said he thought the extra time required of NCEES was worth it because improving teaching practices and teacher leadership was important. This was a much different perception from data collected in the secondary research.

Principals also did not report the time restraints associated with NCEES in the primary study compared to the secondary. Extra supports may have contributed to the difference in perceptions.

Quality of classroom observation and teacher evaluation within teacher evaluation systems. In measuring structured responses on the NCEES Teacher Survey for the category "Quality of Classroom Observations and Teacher Evaluation Within Teacher Evaluation Systems," Items 13 through 25 were designed to measure teacher perceptions of the validity and reliability of ratings across Standards 1-5, the appropriateness of the rubric used to rate teachers across Standards 1-5, and the reliability of observers rating them using Standards 1-5. Item 13 drew a slightly negative response

from teachers at a negative-to-positive ratio of 1.09; while teachers responded positively overall to Items 15, 17, 19, 21, 23, and 25 with a positive-to-negative ratio of 1.83, 2.46, 1.83, 1.34, 2.52, and 1.88, respectively (see Table 8). Unstructured teacher responses measured the effects on teaching practices and teacher leadership for the validity and reliability of Standards 1-5, the appropriateness of the rubric on teaching practices and teacher leadership, and the reliability of observers rating them using Standards 1-5 across Items 16, 18, 20, 22, 24, and 26 (see Appendix A).

Although slightly negative overall, teachers were split almost neutral in their perceptions about the alignment of their ratings on Standards 1-5 with Standard 6 EVAAS® outcomes in structured Item 13. This item measured the validity of teacher ratings for Standards 1-5 by comparing them to teacher EVAAS® outcomes.

The unstructured responses of teachers for Item 14 were as divided as structured Item 13. Some teachers viewed EVAAS® outcomes as a way to know they were increasing the academic achievement of their students, and they viewed their principal ratings from Standards 1-5 as valid in relation to their EVAAS® outcomes. Teachers commented about the positive effects on their teaching practices and teacher leadership, namely the fulfillment that came from knowing they were considered effective teachers using an alternative indicator in conjunction with their principal ratings on Standards 1-5.

A good summary of the positive responses on the unstructured Item 14 involved the overall validation teachers perceived as a result of receiving feedback indicating their effectiveness across multiple measures. The validation effect was vague but implied in the coding process. Rather than straightforward responses, teachers implied there was alignment between their ratings from their principals for Standards 1-5 and Standard 6 EVAAS® outcomes and commented directly about the positive effects. The positive

effects ranged from perceiving their own expertise to devotion in sharing their teaching practices with other teachers.

Specifically, evidence from the codes suggested that teachers who perceived their effectiveness across multiple indicators were likely to hold a desire to improve themselves and others around them. Future studies could collect more data in this area to investigate how "validation" affects teachers who were found to be effective across multiple measures. The "validation" coding reappeared in other items concerning EVAAS® outcomes, but teachers provided sparse details about what it meant and what occurred as a result of "feeling validation."

Other teachers questioned the ability of Standard 6 EVAAS® outcomes to appropriately align with their ratings on Standards 1-5 because their EVAAS® outcomes were based on a single test taken on 1 day throughout a semester of 90 days. When there was misalignment, teachers were concerned with the validity of their EVAAS® rating, not the ratings from their principals for Standards 1-5.

Teachers described the negative effects on teaching practices and teacher leadership when misalignment occurred which ranged from frustration in getting mixed signals about what worked and did not work in their classroom to a general mistrust of EVAAS® outcomes. Some teachers indicated there was a misalignment between their principal ratings for Standards 1-5 and Standard 6 EVAAS® ratings in structured Item 13 and offered input on unstructured Item 14.

- I feel like my ratings for Standards 1-5 are not in alignment with Standard 6. It is a struggle to make them match.
- I feel that my Standard 6 rating has caused my practices to change. I have

changed to more traditional methods of teaching than some of the project based work that I used to do. Personally, I felt that more traditional methods would better equate to improved test scores from EVAAS.

- Quite often my standard 6 scores do not reflect my success in a classroom for that year. It does not show the number of students that I have that are able to complete my course the first time that they take it or those that experience success for the first time since elementary school.

There were also teachers with mixed perceptions. This teacher strongly agreed on structured Item 13 with the alignment of his/her ratings for Standards 1-5 and his/her Standard 6 EVAAS® rating, but also expressed concern in unstructured Item 14:

> My ratings [from my principal] are good and my Standard 6 rating is good. Although I have met or exceeded growth over the last several years, I am more concerned with how my students are performing in my classroom than how they do on one single test given by the state. I work to meet them where they are and for their benefit and their growth as a student and productive citizen; not how they are going to perform on one individual test.

For structured Items 15-25, teachers were positive overall about the ability of NCEES to capture their performance with validity (Items 15, 17, 19, 21) and reliability (Item 23) across a range of standards and about the ability of principals to reliably (Item 25) rate them. This was manifested by high positive-to-negative ratios in their responses. For unstructured Items 16-26, teacher responses paralled their positive responses on the structured Items 15-25.

In considering the ability of NCEES to capture their performance with validity and reliability across a range of standards and the ability of principals to reliably rate

them, teachers provided some evidence for both positive and negative effects on teaching practices and teacher leadership for the unstructured Items 16-26. Because teachers overall trusted the validity and reliability of Standards 1-5, evidence from the coding of their responses suggested teachers have been using standards-based feedback from principals and the rubric to address the learning needs of their students. Specific teacher responses that contributed to this theory included the following across unstructured Items 16-26:

- I use the rubric to improve teaching practices. I want good ratings so I see what actually constitutes one and change accordingly.
- It has confirmed what works well in the classroom and what I need to work on more. This makes my teaching practices stronger and more efficient.
- The rubric is a little long but is a great template describing what I do.
- I agree with the Standards 2-5. This demonstrates exactly what I'm doing in the classroom that has a direct impact on my students.
- The rubric allows me to be able to see weaknesses and prioritize changes in practice.
- I am more aware of opportunities for leadership because of the standard.
- To hit the higher ratings, I strive to lead my students down the path of success. I allow my students to present findings during some projects to show them they have the skills to reach their individual goals. I always receive good feedback from my principal on this. But it is good for my students to teach them how to teach others and be good citizens.
- Encourages to teach in a way mindful of what is looked for on NCEES, and in

every lesson.

- I feel I am better because of the ratings from my administrators show it.
- I want good ratings and good test scores. I make necessary changes to my practices to get these ratings and scores. So far there is a positive impact on myself and my students.
- It has encouraged me to find opportunities to work with my peers to fit the leadership standard. This has helped build collaboration which I believe helps the school as a community and furthers student learning.
- I do try to follow some of the items listed on the rubric, like varying assessments and using technology, so–in a sense–those items have positively affected my teaching practices.
- I feel as though my leadership in education has improved greatly in my ten year career and even more so because of this instrument. I am able to share the things I do with my principal during my summative to show how I have improved.
- It has helped me develop clear goals and understand areas where I need improvement, as well as allowing me to see some strengths I had not noticed before.
- I know I'm doing my job since my ratings are high.
- I pay more attention to standards now than in the past because I worry that I am not doing what my principal wants. Not only do I list my essential standard on my lesson plans, I list standard and element from the instrument by the practices I choose. When it comes to evaluation time I can show my

principal what I'm doing and why.

- Since my ratings are stable from year to year and principal to principal, and my scores have also been good, I figure "if it ain't broke, don't try to fix it." I'll try anything rational as long as my students respond and perform at higher levels.
- A variety of administration and peers helps to be sure I am given a fair evaluation and so far this has happened. I see good results no matter who is evaluating me, and if there is something I hear that needs attention, I fix it.

Although teacher input on unstructured Items 16-26 was positive overall, there were concerns interspersed throughout. These responses centered on unreliable ratings from principals, a perception of the inadequacy of the rubric, and the inability of the rubric and observers to measure teacher leadership. In many cases, these responses also relayed the negative effects on teaching practices and teacher leadership that ensued. A sample of these responses included the following:

- I don't have issues with the rubric . . . just issues with how it's used from time to time depending on who is using it. When I have one person telling me one thing and another something else, what am I supposed to believe?
- I loosely agree. Some subsets of each standard cannot be measured, therefore I disagree with them being a part of my overall evaluation. None of these affect my teaching practices.
- Most of the standards have very little to do with teaching techniques.
- I agree with standards 2-5 but do not know exactly how it will be applied. I feel they need more work outlining what they are looking for and allow

feedback as to why it is important or not important.

- I feel that the rubric is vague for Standards 2 and 5, so I do not really see them as an accurate guide to rate my teaching practices.
- The rubric for these standards includes things that often are not observable, that happen outside of the classroom, requiring teachers to take extra steps to prove the things we do regularly.
- Define leadership prior to using the standard.
- I have mixed feelings on this. Some teachers are not given the opportunities to have leadership through the county office, when others are. I also feel there are many ways you can be a leader in your content area, and administration may not be aware of your efforts or leadership within your department. This gets frustrating for me.
- I feel as though there is a competition for teachers who are not leadership material to be leaders in departments or in small groups just so they can use this as proof on their evaluation. This hurts the collaboration portion of my teaching. Leadership is not just having a title next to your name.
- Teacher leadership involves more than what can be observed in the classroom.
- My daily teaching practices are not adequately addressed with this instrument.
- I am more cognizant now that teacher leadership is something to strive for; however, at my school teacher leadership positions are often appointed. It is difficult to perform well in this category if I am not selected for such responsibilities.

Principals were positive overall about using the NCEES rubric for Standards 1-5

to rate teachers, although there were concerns communicated. There was an overall level of agreement among principals that Standards 1-5 captured a spectrum of teaching practices and teacher leadership that would positively affect teaching practice and teacher leadership when teachers were rated accurately over time.

Although principals were positive overall about Standards 1-5 of NCEES, there were several concerns expressed by multiple principals surrounding Standard 4. The first and foremost concern centered on a lack of an element in the NCEES rubric necessary to gauge the classroom management skills of teachers. Multiple principals communicated that they felt teaching practices were negatively influenced by NCEES not containing a specific element within Standards 1 and 4 to rate teachers on how well they managed potential discipline issues. Multiple principals argued that the existing elements in Standards 1, 2 and 4 (1a, 2a, & 4b) were inadequate and vague in rating how well teachers managed student behavior, which principals described as an important issue especially facing beginning teachers.

Along with the critique by principals of Standard 4 missing specificity in giving principals the ability to rate a teacher in their efforts to maintain student discipline, one principal criticized Element F in Standard 4 where teachers have been rated on their ability to form student-led teams to promote student learning. The principal thought it might not be possible with some at-risk students to design such a setup.

A principal also advocated a shorter version of the current rubric used within NCEES for beginning teachers. There was a concern that the current rubric was too long and sophisticated for beginning teachers to understand and fulfill. This principal believed such a long, sophisticated process might have detrimental effects on the teaching practices of beginning teachers.

In summary, teachers responded positively overall to both the structured Items 15, 17, 19, 21, 23, and 25 and their connected unstructured Items 16, 18, 20, 22, 24, and 26. Item 13 received an overall negative response in both the structured and unstructured portions and the negative responses centered on the inadequacy of Standard 6 EVAAS® outcomes to capture their teaching practices in a valid way. In turn, the negative effects on teaching practices was defined by a mistrust of Standard 6 EVAAS® outcomes.

In contrast, teachers who responded positively to Item 13 did so under the pretense of experiencing alignment between their ratings for Standards 1-5 and Standard 6 EVAAS® outcomes. These teachers communicated a code entitled "Validation Effect"; where they felt validated, they were effective teachers and desired to be innovate to improve their own teaching practices and those around them.

For unstructured Items 16-26, depending upon whether teachers were positive and negative about the different factors involving the quality of NCEES, they revealed varying effects on teaching practices and teacher leadership. The positive effects on teaching practices and teacher leadership included internalizing the rubric in an attempt to achieve higher ratings from their principals, the willingness to make changes to improve teaching practices and teacher leadership based on the rubric and feedback from principals, and the ultimate outcome of improved student achievement by improving teaching practices. The negative effects on teaching practices and teacher leadership included the unwillingness to change teaching practices and teacher leadership due to unreliable ratings from observers, the vagueness of the rubric, and confusion and unfair practices over what constitutes teacher leadership.

Principals expressed an overall acceptance of Standards 1-5 within the rubric used with NCEES but communicated concerns about a lack of an ability to rate teachers based

on their ability to diffuse discipline issues. Principals also were skeptical of Standard 4 Element F because of the inability of teachers with at-risk students to carry out student-led teams. A principal also expressed possible frustration for beginning teachers in using the current rubric because of its length and level of sophistication.

Use of VAM within teacher evaluation systems. In the category "Use of Value-Added Modeling Within Teacher Evaluation Systems," structured Items 27, 29, 31, and 33 were dedicated to measuring teacher perceptions of the validity, reliability, and unintended consequences of EVAAS® ratings; while unstructured Items 28, 30, 32, and 34 measured the effects on teaching practices and teacher leadership in free-response boxes (see Appendix A). Structured Items 27, 29, and 31 experienced overall negative responses with negative-to-positive ratios of 1.52, 1.59, and 1.03, respectively; while Item 33 was slightly positive with a positive-to-negative ratio of 1.11 (see Table 8).

The trend that began with teacher perceptions of Standard 6 EVAAS® ratings in Item 14 continued throughout items in this category. Teachers who perceived EVAAS® outcomes with a positive tone in the unstructured portions of EVAAS®-related questions and agreed or strongly agreed in the structured portions viewed EVAAS® outcomes as validation of being a competent teacher. This was in important idea interwoven through memo writing and eventual theory building around EVAAS®-related questions–teachers appreciated another line of positive feedback about their teaching practices and EVAAS® provided it.

Another idea that migrated upwards within the coding process for this category was that teachers who were positive in their perceptions of EVAAS® outcomes had forward-looking perspectives, and the forward-looking perspectives were an extension of their validation. Teachers with positive perceptions of EVAAS® used past successes with

students within the purview of EVAAS®, validated themselves as competent teachers, and were positive about their teaching practices with future students as a result. There were teacher comments that were indicative of this general message from teachers with positive perceptions of EVAAS®, of which included the following:

- I was very delighted in the fall when my students' scores reflected that they had grown. I felt relieved as a first year teacher that my students were gaining stronger abilities in Language Arts and feel confident about my kids coming in next year based on these results.
- After the first EVAAS scores were released I was able to gain knowledge on how to assist my future students in order to help them grow.
- I already knew I was a good teacher from input I get from students, peers, and past and present admins. I know EVAAS has reinforced this.
- I use EVAAS as a basis for what I do from year to year. If the practices I use are successful this year, I will concentrate on using those same practices from year to year.

Teachers with positive perceptions of EVAAS® on the structured items did not choose to characterize the types of students they taught in the unstructured items as behaviorally challenged, at-risk, bad test takers, nonhonors, exceptional children, special needs, absentees, potential dropouts, below grade level, or poor readers. Conversely, teachers with negative perceptions of EVAAS® on the structured items used all of these labels across unstructured responses to describe the types of students they taught to explain why EVAAS® outcomes have not provided a rating that reflected the academic growth of their students (Item 27) and to explain why EVAAS® outcomes have not

reflected the quality of their teaching practices (Item 29). Teachers with negative perceptions of EVAAS® outcomes oftentimes expressed that they were better judges of the academic growth of the various types of learners they faced compared to student growth based on standardized tests used to calculate EVAAS® outcomes. Examples of these comments were

- I struggle with finding ways to help my students who have difficulty decoding NCFE questions and are “bad test takers.”
- Honors and growth do not go well together. They do not correlate well with what they actually learned.
- Doesn't account for student abilities, limitations, absences, etc. It’s all or nothing!
- I believe this test reflects the reading and test-taking ability of each of my students on a given day. I do not necessarily agree that this ONE test accurately reflects each of my student's true abilities. I believe some did his or her best, but due to a variety of circumstances will never be considered PROFICIENT. Others, despite a teacher's exceptional teaching ability, will do well regardless. I know that I have taught to the best of my ability, every student that has entered my classroom, the very intelligent to the reluctant learner to the ELL.
- I implement various projects and labs in my instruction. My students enjoy this teaching practice, and they have proven to me that they know the material. However, some of them get confused on multiple-choice question strategies. Just because they get 1 or 2 questions wrong on their NCFE about a

particular standard, it doesn't mean they don't know anything from that standard. I struggle with finding strategies to, "teach to the test" like I'm supposed to.

- I work very hard all year to help my students, but unfortunately, many of them are far below grade level or do not care about exams, so I do not feel that this affects my practices other than to make me discouraged.
- I just disagree because I teach behaviorally challenged student. Majority of my classes are inclusion EC with no real help.
- Since I work with a special needs population that does not always show a lot of growth, I don't believe Standard 6 is an accurate reflection of my teaching practices.
- I believe that it is hard to get an accurate reading of my quality of teaching when I have a few students who have missed numerous days of classroom instruction. It is very hard for a student who has missed 20+ days of instruction to do well on the North Carolina Final Exam. Does EVAAS take this into account?
- I have been told to keep on doing what I have always done in the classroom. I follow the three R's, which are rigor, relevance, and relationships. I don't think that the ratings are very reliable when there are so many outside societal forces that are negatively affecting the lives of my students.
- I received good ratings, yet I think they are not always indicative of every student. Some are not good test takers and others may have missed substantial pieces of content during an extended illness or tragedy in their life.

Teacher responses to unstructured EVAAS®-related items were as polar as the EVAAS® category in the literature review. As with the literature review, teachers with positive perceptions made arguments for its use; and those with negative perceptions made arguments against its use. The vast majority of teachers who were neutral in their perceptions on structured items involving EVAAS® supplied no responses to the unstructured items involving EVAAS® outcomes.

Principals overall did not concur with the perspectives of teachers on the basis of viewing Standard 6 EVAAS® outcomes through the lens of student characteristics, but there was one important exception. The principal at a small alternative high school communicated skepticism in the use of EVAAS® given the demographics of the students at this school. This principal did mention that he has witnessed some unreliability and jumps in EVAAS® outcomes for certain teachers.

The principal communicated many issues that he saw as controversial in using EVAAS® outcomes to rate teacher performance in this alternative school. Although the principal appreciated the diagnostic value of EVAAS®, the principal expressed the following concerns surrounding EVAAS® outcomes and the types of students he had enrolled at the alternative school:

> It goes back to the unpredictable performance of our kids . . . they're in and out. We have students transition in and out with quick turn over in some cases. We have little control over where we get them. Some come from schools, some come from other places like lock down facilities. They'll be here for a week. So it's like a revolving door for some of these kids. There is some reservation for us in alternative education compared to traditional schools. I could take five of our students at random and place them in a high school somewhere and they would

> completely disrupt the environment of that school. We have to deal with many more hardships . . . I'm talking extreme hardships like dealing with kids on probation, and I'm talking about kids involved in some heavy duty crime. Is it fair for our teachers to be judged on the basis of the test scores of these students? Then we've had a running debate with our schools and with decision makers at the state level. What cut-off date should we use for our teachers to claim students for EVAAS? When they're in and out, what's a fair cut off for teachers before we just say to not count them at all?

Another important outcome of the coding of Standard 6 EVAAS®-related items was the straightforward admission of teachers about their lack of knowledge of how EVAAS® worked to calculate student growth. Teachers' lack of knowledge showed up in codes such as "I do not understand," "No one can explain why," "show/explain to me," "I don't believe I've had enough experience," "I have no clue," "Nobody knows anything," "not sure," "is so abstract," "I am unsure," "I have no idea why," and "I don't see." Examples were as follows:

- I am not sure how EVAAS works. I have not changed a thing but I went from growth to average. No one can explain why. My principal says I am doing a good job and wait to see what happens next year.
- This is a subjective piece that I am unsure about in terms of inter-workings. To me, I have little control over how my kids perform on their tests.
- I do not understand EVAAS in the least, and by the time I do receive the data, my kids have already moved on from my class, so it does not affect my teaching practices and leadership in the least.

- I wish someone who show/explain EVAAS data usage to me.
- EVAAS is so abstract and confusing that it provides absolutely NO input to affect my teaching practices and leadership.

These codes led to teachers revealing the negative effects of EVAAS® outcomes on teaching practices by producing codes with negative processes such as "we do not look at EVAAS," "I can't comment," "It's discouraging," "Frustration," and "it is not fair." Examples of outcomes negatively affecting teaching practices were as follows:

- I honestly don't know what my Standard 6 rating is. I check my test scores and student data on EVAAS, and that's as far as I go.
- I have had different results while no one can specifically point me in the right direction. I feel like this is something I have no control over.
- Considering that my course takes the North Carolina final exam and I have no clue how the scores are calculated nor whether the free responses are even graded and calculated within the scores I can't comment on whether it's effective or not.
- It's discouraging how teachers are evaluated on student academic growth feel set up for failure.
- Frustration because all admin sees is numbers not the whole class from dynamic picture.
- I work extremely hard and standard 6 could never fairly evaluate it.

In interviews, principals were aware of negative teacher perceptions of using EVAAS® within NCEES; and in some cases, principals mentioned that teachers lacked a level of knowledge to understand how EVAAS® worked to calculate student growth.

Principals also attested to correcting teacher misconceptions about Standard 6 EVAAS®, although one assistant principal replied that they were not as supportive in using EVAAS® outcomes during the past school year. In this particular school, there were changes in administration; and this assistant principal was confident that Standard 6 EVAAS® outcomes would be a more integral part of their approach in upcoming years.

There were two teacher misconceptions that principals widely stated they attempted to correct about Standard 6 EVAAS® ratings. First, principals overall have attempted to curtail negative perceptions of teachers in using the characteristics of their students as the reasoning for the negative perceptions surrounding EVAAS®. The message to teachers has been that student characteristics have been controlled for within EVAAS® calculations. These principals also stated that teachers have received EVAAS® training in past years through curriculum coaches who have informed teachers that the characteristics of students should not be a factor over many years of data–specifically that students act as their own controls and that their past performance across many years of testing should be reflected in their present performance.

Second, principals have tried overall to inform teachers that EVAAS® ratings have not been based on one test on 1 day. Principals have actively tried to curtail this belief in teachers both in personal and faculty-wide discussions, alongside training from curriculum coaches. According to principals, teachers have been informed that Standard 6 EVAAS® ratings have been compiled over many years of data in order to calculate a conservative teacher effect on student growth. Principals realized overall that calculating EVAAS® ratings using one, two, or three classes in a semester would not produce an accurate outcome and that EVAAS® ratings have been generated using testing data over many years which includes thousands of testing data points for hundreds of students for

each teacher. Principals overall also said they have relayed these talking points to their teachers during discussions.

The negative effects of teacher misunderstandings within EVAAS®-related items were still apparent regardless of the steps taken at the school-level to counter them. When teachers carried negative perceptions of EVAAS® outcomes, they expressed frustration and an unwillingness to use EVAAS® as a diagnostic tool to improve student learning. While teachers carrying negative perceptions of EVAAS® outcomes communicated the negative effects on their teaching practice, teachers with positive perceptions included details of how EVAAS® outcomes have positively affected their teaching practices. Responses such as these characterized the positive responses:

- I noticed that my weaker students met growth, but my stronger students were often below their expected growth. This has led me to reflect on whether I am challenging my higher performing students.
- Standard 6 has made me work harder than ever to help my students learn better test-taking skills. It is hard to teach a 9th grader reading on a 2nd grade level to read a poem or fictional passage written on a 10th grade level. The best I can do is to teach the concepts, critical thinking, and test-taking skills. I have done more tutoring than ever (with no more incentive), as have the students. They want to be successful, and I want them to be. I have worked as hard as possible to help my students through instruction, additional tutoring and test-prep to prepare for the NC Final Exam. I hope that it pays off.

Teachers sent a clear message throughout all EVAAS®-related items on the NCEES Teacher Survey that they were concerned about the phenomenon of teaching to

the test. Item 33 was designed to measure teacher perceptions of the influence of Standard 6 EVAAS® on teaching practices; and although Item 33 received a slightly positive-to-negative ratio of 1.11, teaching to the test was a prevalent coded response in the follow-up unstructured Item 34.

One connecting attribute of codes across unstructured Item 34 was the insight from teachers that they had already been searching for ways to increase student achievement before the implementation of EVAAS®. A memo named "In Conjunction" became apparent where teachers felt as though EVAAS® added to their already existing search to find innovative teaching practices to boost student achievement.

Teachers cited their responsibilities as professionals as the driving factor for searching out ways to increase student achievement. Structured Item 33 was the only structured item receiving an overall positive-to-negative ratio that measured teacher perceptions about EVAAS®-related constructs, although the positive-to-negative ratio was small. This item served as a minor contradiction in the outcomes relating to EVAAS®-related items.

As a result of this minor contradiction, teachers who "disagreed" and "strongly disagreed" on EVAAS®-related items were tracked to determine if and why they changed their perceptions, and coding was compared to discover perceptions to Item 33 before responding and after. Evidence in the codes suggested that the increase in positive perceptions about EVAAS® in Item 33 was connected to the appreciation they had for already seeking out ways to be innovative in their teaching practices, how teachers measured the outcomes of those innovations, and the mindful acceptance of welcoming another stimulus as a driver.

Some comments that reflected the connection between what teachers were already

doing before the addition of EVAAS® and the general acceptance of EVAAS® as a driver adding to the already existing motivation to seek and use teaching practices that boosted student achievement were as follows:

- I would seek and use, teaching practices that help my students grow academically, with or without Standard 6.
- I look for new practices to help me and my students grow regardless of my rating.
- I did this before EVAAS and Standard 6 and I will continue.
- Standard 6 has influenced me to seek and use best teaching practices each and every year. I have always and will continue to do that because I am a professional who values her students' learning.
- I want to seek and use new and better teaching practices because that is the right thing to do, not for ratings. EVAAS has not changed the way I do things because I was already doing them.
- I am always looking for new ways and strategies to enhance my teaching practices and I continually strive to help my students grow academically.
- Although I do not use NCEES to directly determine my teaching practices, I seek and use teaching practices that I believe will help my students grow because I have been trained to do so in college and because I am a professional who makes these judgments based on my education and on previous experience.

One unintended consequence of implementing VAM in the literature review was the idea that VAM outcomes were driving principals to inflate or deflate their ratings to

follow suit with VAM outcomes (Amrein-Beardsley & Collins, 2012; Headden, 2011). While principals were aware of specific EVAAS® scores for their teachers, there was no evidence from principal interviews or teacher responses on the unstructured items that EVAAS® ratings in Standard 6 were driving principals to rate teachers differently on Standards 1-5.

Many principals were unable to provide a perception of whether their ratings for Standards 1-5 were in line with specific EVAAS® scores for their teachers. Principals concentrated more on using EVAAS® outcomes to provide support for struggling teachers and to improve teachers along the spectrum of performance.

This provided indirect evidence that EVAAS® outcomes were not driving principals to change their ratings on Standards 1-5 to match EVAAS® outcomes; because if they suggested they were cognizant of the alignment, it could imply they were concerned. A more direct mode of evidence was that the NCTSR movement across years showed principals rating teachers with less and less "developing" and "distinguished" ratings across years, while EVAAS® ratings were consistently providing higher percentage of ratings for teachers at the levels of "does not meet expected growth" and "exceeds expected growth." If principals were ramping up (or down) their ratings to match EVAAS® outcomes, the movement of NCTSR from year to year would not have been being compressed towards the middle ratings away from "developed" and "distinguished" ratings; rather, the NCTSR movement across years would have been expanding outwards towards the "developing" and "distinguished" ratings.

This was an important outcome of the primary study because if principals were deflating or inflating their ratings to match EVAAS® outcomes, there could be negative effects on teaching practices and teacher leadership in that teacher performance would be

solely judged based on a single indicator. Compounding the problem would be that EVAAS® and VAM outcomes in general have been shown to be unstable across years; which would mean if principals were matching their ratings with EVAAS® outcomes, instability would ensue with large spikes in ratings for some teachers, and those particular teachers could become confused as to their effectiveness. Given the two lines of aforementioned evidence, principals have shown to be unbiased at least by EVAAS® outcomes when rating teachers in the county under study.

The positive effects of experiencing principal ratings independent of EVAAS® outcomes could be inferred to involve principals providing the truest judgment of teacher performance possible. This would provide teachers with accurate feedback as to how they could improve their teaching practices. Also, when principals rate teachers independent of EVAAS® outcomes, they act as a safeguard against false positives or false negatives shown to occur within VAM as Glazerman et al. (2010) and Schochet and Chiang (2010) found.

In summary, teachers perceived EVAAS® outcomes in a slightly negative sense within the structured items and concentrated their negativity on the validity and reliability of Standard 6 and EVAAS® in the unstructured items. Teachers questioned the ability of EVAAS® to measure the quality of their teaching practices and the academic growth of their students because teachers viewed the characteristics of their students as an overriding factor. Although principals have attempted to thwart teacher misconceptions surrounding Standard 6 and EVAAS® through personal and faculty-wide discussions and training within PLCs, some teachers continue to cite and uphold what the literature has deemed misconceptions as justification for the lack of validity in Standard 6 EVAAS®.

The positive effects of Standard 6 EVAAS® ratings were that teachers who

perceived EVAAS® in a positive light viewed EVAAS® outcomes as validation of competency and were motivated with a future perspective to increase student achievement as a result. The negative effects on teaching practices were that teachers dismissed EVAAS® outcomes or chose to ignore EVAAS® outcomes using justification that EVAAS® was not valid or reliable. Teachers who were negative toward EVAAS® outcomes also chose to label their students based on their characteristics without evidence of an attempt to overcome student characteristics to increase their academic achievement.

Structured Item 33 served as a minor contradiction in EVAAS®-related constructs. While other items regarding EVAAS® incited negativity, responses for Item 33 garnered a slight positive-to-negative ratio. After investigation, it was determined that there was a level of appreciation from teachers who previously viewed EVAAS® with negativity for adding to their already existing motivation to find teaching practices that boost student achievement.

It was also determined from two lines of evidence that EVAAS® was not influencing principals to rate teachers any differently than based on their own judgments and based on the performance of teachers they observed in the classroom. The positive effect of this finding was that such unbiased classroom observations has led to more concrete feedback to teachers in improving student achievement.

How are outcomes from teacher evaluation systems utilized? In measuring structured responses on the NCEES Teacher Survey for the category "How Are Outcomes from Teacher Evaluation Systems Utilized," Items 35 through 41 were designed to measure teacher perceptions of the usefulness of feedback from principals for improvement of teaching practices and teacher leadership within times of formative and summative evaluations. Items 35 and 39 drew overall negative responses from teachers at

negative-to-positive ratios of 1.31 and 1.73, respectively, and measured teacher perceptions of their ability to improve teaching practices during formative (Item 35) and summative feedback (Item 39); while Items 37 and 41 drew overall positive responses at positive-to-negative ratios of 2.79 and 1.58 and measured teacher perceptions of their ability to improve teacher leadership during formative (Item 37) and summative feedback (Item 41) (see Table 8). Unstructured Items 36-42 involved teacher free responses about the usefulness of feedback from principals for improvement of teaching practices and teacher leadership within formative and summative evaluation (see Appendix A).

The results of structured Items 35 and 39 compared to structured Items 37 and 41 presented the highest degree of polarity among items on the NCEES Teacher Survey. To discover the cause of this polarity, an investigation ensued that paralled the investigation in the previous category ("Use of Value-Added Modeling Within Teacher Evaluation Systems") where Item 33 was positively perceived by teachers and diverged from the negatively perceived items related to EVAAS® (Items 27, 29, and 31).

Teacher responses to structured Items 35-41 and unstructured Items 36-42 were tracked by specific teachers in an attempt to discover why teachers had changed in their perceptions on various items. Teachers with initial negative responses to structured items 35 and 39 were tracked for comparison in attempting to reveal why they changed their perceptions in structured Items 37 and 41. This was accomplished by an analysis of responses of tracked teachers in unstructured Items 36 and 38 to uncover differences in perceptions in structured Items 35 and 37 and an analysis of responses of tracked teachers in unstructured Items 40 and 42 to uncover differences in perceptions in structured Items 39 and 41.

Principal and teacher perceptions on unstructured Item 35 were already revealed

in a previous section entitled “Synthesis of Novel Categories and Conceptual Questions Using Grounded Theory” and examined in order to build the novel category “Principal and Teacher Expectations of Feedback.” Teachers revealed that there was not enough feedback; the feedback was not as specific as teachers desired; and the feedback did not match expectations teachers held for their particular subject area.

However, there were important differences uncovered between Items 35 and 37 based on tracking the changes in perceptions for specific teachers who were initially positive in structured Item 35 only to change their perception to negative in structured Item 37. This subsample of teachers under analysis revealed that teachers who perceived they had not received enough feedback or that the feedback was not specific enough to positively affect their teaching practices in structured Item 35 changed their perceptions in structured Item 37 because there was an adequate amount of specific feedback offered to improve their teacher leadership. This subsample of teachers were observed “straddling” the “neither agree nor disagree” “fence” in Table 7 and their change in perception was documented in the unstructured Items 36 and 38.

Of the subsample of teachers “straddling,” a smaller number offered responses on unstructured Items 36 and 38. Evidence suggested that this subsample of teachers was positive about the feedback that principals offered them regarding their leadership roles because of their unstructured responses on Item 38. To reiterate, these unstructured responses were from teachers who initially held negative perceptions in structured Item 35, changed their perception in structured Item 37, and offered an unstructured response in Item 38. The specific responses were as follows:

- I have tried to serve on more committees and in more leadership roles this school year.

- I continue to try to lead others in my areas of strength.
- I have taken on leadership positions within my school such as department chairperson and department rep at SIT meetings. I have also participated in various committees.
- I have become more involved in district leadership and school leadership roles. I am a member of committees at the district level and my school's SIT.

Although these responses represent a small sample of responses from the initial subsample of teachers, they supported a theory that principals have been offering more specific feedback as to how teachers could improve their teacher leadership roles during times of formative feedback compared to that of how teachers could improve teaching practices.

Most importantly, codes tended to point to a reason for teachers being more positively inclined to accept principal feedback to improve teacher leadership. That reason involved the specificity in which principals seemed to offer feedback to improve teacher leadership. A memo entitled "Simplicity in Feedback" grew from memos noting the simplistic type of feedback contained in improving teacher leadership compared to that of improving teaching practices. Teachers mentioned specific activities in which principals provided feedback in which teachers were motivated to participate in order to achieve higher ratings from principals. The simplistic feedback offered by principals to teachers mostly involved teachers engaging in collaborative efforts in PLCs or other forms of committees or groups.

In contrast, teachers failed to include any substantial evidence that specific teaching practices were being included in feedback from principals. Teachers with

negative perceptions in general and across categories in the NCEES Teacher Survey often responded with an explanation for why there has been little change in their teaching practices and teacher leadership, and that explanation included the idea that their classrooms and students were too complex for anyone (principal) or anything (EVAAS®) to judge. Teachers responded similarly in terms of using formative and summative feedback to improve teaching practices compared to using formative and summative feedback to improve teacher leadership–teacher leadership also involved somewhat complex tasks. However, there was evidence that principals used specific formative and summative feedback to overcome some of the complexity surrounding what entailed the definition and function of teacher leadership.

Although principals offering specific feedback to improve teacher leadership have yielded positive effects on teacher leadership by giving teachers the opportunity to lead in specific areas, the lack of specific feedback to improve teaching practices has negatively affected teaching practices. The negative effects on teaching practices were apparent when teachers were unable to recount specific ways in which they improved their teaching practices, and in turn there lacked evidence that student achievement improved as a result.

Conversely, the positive effects of teachers improving their leadership roles were revealed as teachers sought to collaborate with and professionally develop their colleagues. Coded data across unstructured items involving formative and summative feedback to improve teacher leadership shared a distinct memo–teachers sought to accept leadership positions within PLCs in an attempt to improve the teaching practices of colleagues ultimately to boost student achievement.

A similar cross analysis for Items 39 and 41 took place to investigate shifts

between the different perceptions of teachers concerning improvements they made in their teacher leadership (Item 41) compared to their teaching practices (Item 39) during times of summative feedback. Unlike the differences detected between structured Items 35 and 37, there was an inability to detect the reason for the shifts between Items 39 and 41 for three reasons.

First, there were 16 responses to unstructured responses for Item 42 and 18 for Item 40 limiting the amount of data under analysis. To put this in perspective, there were 58 responses to unstructured Item 8, which was the first unstructured response available to teachers at the beginning of the survey. Second, teachers who changed perceptions did not respond readily on unstructured Item 42 or their responses were irrelevant to this particular investigation. Finally, the difference in the negative-to-positive ratio of Item 39 (1.73) and the positive-to-negative ratio of Item 41 (1.58) were not as extreme as the negative-to-positive ratio of Item 35 (1.31) and the positive-to-negative ratio of Item 37 (2.79), which meant teachers probably would not display perceptions that were as extreme in unstructured Items 40 and 42.

However, there were unstructured responses from teachers in Item 40 who held negative perceptions regarding the feedback they received to improve their teaching practices during times of summative evaluation (Item 39). These responses were as follows:

- The administrator that did my summative had not even been in my room to observe me this year. How can he give me feedback?
- I just sign.
- Too brief and like an assembly line with people scheduled directly before and

after me.

- I have asked during my summative conference what I should focus on for improving next year and have been told nothing specific, and with the standards being very broad it is hard to point to something finite. I focus more on what I have noticed than what NCEES shows.

These responses from teachers indicated that similar to unstructured Item 36 where teachers perceived they were not receiving enough or specific feedback to improve teaching practices during times of formative evaluation, teachers were also not receiving enough or specific feedback to improve teaching practices during summative evaluation observed in unstructured Item 40. The negative effects on teaching practices were apparent in that teachers perceived a lack in ability to improve their teaching practices because they lacked the guidance they desired.

Although structured Item 41 drew a positive-to-negative ratio in measuring the ability of teachers to improve their teacher leadership during times of summative feedback, it did so with a smaller ratio (1.58) compared to that of structured Item 37 (2.79) where teachers recounted specific ways in which their teacher leadership improved in times of formative feedback. There were responses on unstructured Item 42 that revealed negative effects on teacher leadership when applicable feedback was not offered during times of summative evaluation. These responses were as follows:

- I have tried to get on the SIT for the last two of the 3 years to improve this one standard, but I have not been “elected” yet.
- The only thing I worry about with the summative is not to be marked “developing.” Since I have never been marked “developing,” I don't worry

about the summative, and it doesn't affect my teacher leadership.

Two theories emerged as to why teachers had more positive perceptions regarding feedback during times of formative feedback compared to times of summative feedback. First, times of summative evaluation consisting of pre and postconferences were being rushed with comments such as “I just sign” and they were “too brief and like an assembly line.” Second, principals carrying out summative pre and postconferences had not observed teachers in their classrooms but were rather sharing final ratings recorded online from other observers.

The positive effects on teacher leadership during times of summative evaluation reflected similar positive effects observed during times of formative evaluation. In both cases, codes at the lowest levels of analysis shared the same common thread. In both formative and summative evaluation, teachers have been driven to seek out and accept positions of leadership where they can positively contribute to professionally developing their peers.

In summary, the teacher reponses to Items 35 and 37 alongside Items 39 and 41 presented the highest degree of polarity–teachers tended to straddle the “neither agree nor disagree” response opting for “agree” in Items 37 and 41 and for “disagree” in Items 35 and 39. Because of this polarity, an investigation ensued that involved tracking teachers who “straddled” and switched perceptions from item to item.

The ensuing investigation revealed evidence that teachers who initially responded negatively to receiving useful feedback to improve their teaching practices in structured Item 35 responded positively to receiving useful feedback to improve their teacher leadership in structured Item 37. Unstructured Item 36 revealed that teachers were unsatisfied with the amount and specificity of feedback offered to them to improve their

teaching practices, whereas unstructured Item 38 revealed that teachers were satisfied with the specific nature of feedback that was being offered to improve their teacher leadership.

The investigation of polarity between structured Items 39 and 41 resulted in a lack of evidence to explain the switch in perceptions. There were three reasons for this–there was not enough data collected in unstructured Items 40 and 42; teachers who switched perceptions did not readily respond across unstructured Items 40 and 42; and the difference in the negative-to-positive ratio of Item 39 and positive-to-negative ratio of Item 41 was not as extreme.

The positive and negative effects of the usefulness of formative and summative evaluation to improve teaching practices and teacher leadership were noted. Teachers who carried negative perceptions of improving teaching practices based on feedback from principals in times of both formative and summative feedback perceived a lack in the amount and specificity in which feedback was offered. This was especially true in times of summative evaluation where teachers perceived that they were rushed through the pre and postconferences and where teachers perceived a lack of feedback from the principal carrying out the summative evaluation because the principal did not observe the teacher in the classroom throughout the year. Teachers also carried negative perceptions of improving teacher leadership when they perceived that they were not electable to leadership positions, in turn offering them no option to improve others around them.

The positive effects of the usefulness of formative and summative evaluation to improve teaching practices and teacher leadership shared similarities throughout the coding process from the lower to higher levels. Teachers who perceived that they had received useful feedback from principals were highly motivated to seek out and accept

positions of leadership where they could professionally develop their peers to improve teaching practices on the aggregate and boost student achievement.

Summary of teacher responses on the NCEES Teacher Survey. In summary, teachers participating in the NCEES Teacher Survey were positive overall about whether they received an adequate number of formal (Item 7) and informal classroom observations (Item 9) and about the amount of time they had to participate in pre and postconferences (Item 11).

Following structured Items 7, 9, and 11, unstructured Items 8, 10, and 12 were free-response questions and measured teacher perceptions of how the number of formal (Item 8) and informal classroom observations (Item 10) and the time they had to participate in pre and postconferences (Item 12) affected their teaching practices and teacher leadership.

In unstructured Items 8 and 10, teachers looked more favorably upon formal classroom observations than informal because of the amount of data collected in formal observations, which could then be used to improve teaching practices and teacher leadership. Teachers also looked at classroom observations as an accountability mechanism ensuring they and their peers were using quality teaching practices and teacher leadership. In unstructured Item 12, teachers did not respond that they were overly concerned with time restraints, a theme found in the pilot data and literature review (Bryant, 2013).

Teachers were slightly negative overall about the validity of Standards 1-5 in relation to their EVAAS® outcomes (Item 13) but positive overall about the appropriateness of the rubric for Standards 1-5 (Items 15 and 17), the validity and reliability of Standards 1-5 (Items 19, 21, and 23), and the reliability of observers rating

them using Standards 1-5 (Item 25).

Unstructured Item 13 initiated a theme redundant in structured and unstructured items following that asked teachers about EVAAS® outcomes. Teachers were slightly negative overall about their alignment of Standards 1-5 with Standard 6 EVAAS® outcomes, and teachers relayed negative effects of which "teaching to the test" was the most redundant and continued to be in future EVAAS®-related items. Teachers with positive perceptions communicated validation in that they were considered effective across multiple measures.

In unstructured Items 14 through 26, teachers were positive overall in their responses with an appreciation for the dimension of Standards 1-5 in measuring teacher performance that improved the academic achievement of their students. Teachers internalized the rubric and, because of the credibility they gave the rubric used to rate teachers, integrated Standards 1-5 into their lesson planning.

Teachers were slightly negative overall in their perceptions of the validity and reliability of EVAAS® outcomes but were slightly positive overall in perceiving a lack of unintended consequences associated with EVAAS® outcomes. Teachers responded overall that teaching to the test was an issue plaguing their approach. Also, teachers with negative perceptions of EVAAS® continually cited the complexity of their classrooms and the characteristics of their students as reasons why they considered EVAAS® outcomes invalid or unreliable.

Teachers presented different perceptions of how outcomes were utilized within NCEES. Teachers were negative overall about the usefulness of formative feedback from principals to improve teaching practices. In contrast, teachers were positive overall about the usefulness of formative feedback from principals to improve teacher leadership. The

same contrasting outcome occurred between the usefulness of summative feedback from principals to improve teaching practices and the usefulness of summative feedback from principals to improve teacher leadership–teachers were negative overall about summative feedback to improve teaching practices but were positive overall about summative feedback to improve teacher leadership.

This contrast led to an investigation that tracked teacher responses across items based on whether they had initially responded negative across items involving the usefulness of formative or summative feedback to improve teaching practices and followed by responding positively across items involving the usefulness of formative or summative feedback to improve teacher leadership. It was determined that teachers responded more favorably to receiving feedback pertaining to leadership than to teaching practices because of the simplicity and specificity incorporated within feedback to improve leadership compared to the feedback to improve teaching practices.

Inferential statistical tests showed significant differences between certain demographic groups and their responses on the NCEES Teacher Survey. Teachers with more years of experienced responded significantly different than teachers with less experience about the time they had to participate in pre and postconferences. It was determined that teachers with less years of experience agreed proportionally that they had adequate time to participate in pre and postconferences at a higher rate proportionally than did teachers with less experience.

Teachers with more years of experience and teachers who had experienced a previous teacher evaluation system agreed at a high rate proportionally compared to teachers with less years of experience and teachers who had not experienced a previous teacher evaluation system that Standard 6 EVAAS® outcomes provided a rating that

reflected the academic growth of their students.

Teachers with less years of experience and teachers who had not experienced a previous teacher evaluation system agreed at a high rate proportionally compared to teachers with less experience and teachers who had not experienced a previous teacher evaluation system that they received useful feedback from observers to improve their teaching practices.

Principals provided qualitative data that overall paralleled teacher input on unstructured items on the NCEES Teacher Survey. The only exception surrounded the use of EVAAS® where principals were attempting to provide solutions to teacher misconceptions. Evidence from teacher responses suggested teachers were not accepting the solutions principals were offering and misconceptions still existed.

Summary of principal perceptions. Because of the open-ended nature of the items used within the NCEES Principal Interview, principals communicated along a wide response range. Principals provided evidence mostly of the positive effects of NCEES on teaching practices and teacher leadership with a few notable negative effects of NCEES on teaching practices.

Principals designed useful ways to increase the number of formal and informal classroom observations. Principals took advantage of outside observers to collaborate in determining teacher ratings. For instance, four of the six principals interviewed were able to take advantage of outside observers because of varying resources. One principal was able to utilize outside observers from the county office because of the physical location of his campus; another principal was afforded an extra assistant principal due to an internship; and another principal noted that there was an extra assistant principal allotted compared to other schools based on their sizes. A different principal also used the ninth-

grade academy coordinator who was appropriately certified to participate in classroom observations.

In all of these cases, principals made use of these extra positions to increase the number of classroom observations. It was also inferred that because of these extra resources, the time necessary for principals to operate was reduced. Principals overall expressed that NCEES required extended time, but the data these principals supplied were in sharp contrast with data from the literature and pilot data where principals in Tennessee (TNDOE, 2012) and principals in pilot schools heavily focused on a shortage of time.

Principals identified that by observing teachers consistently and their teaching practices and teacher leadership roles, they have been able to provide a higher level of feedback. Principals uniformly agreed that their feedback was more valid and reliable as a result of increased formal and informal classroom observations and the collaboration among principals (and assistant principals) to provide teachers with ways to improve their teaching practices and teacher leadership. Principals often used PLCs to provide a mechanism in which they could implement wide-scale improvements that resulted from the validity and reliability of formal and informal classroom observations.

Principals were overall accepting of the NCEES rubric and Standards 1-5, but there were a few notable exceptions. Principals appreciated the spectrum of teaching practices and teacher leadership represented in the rubric used within NCEES and overall agreed that the rubric has the potential to initiate discussions with teachers who would ultimately improve teaching practices and teacher leadership along with student achievement.

Principals provided notable exceptions to their overall positive position on the

NCEES standards and rubric. First, principals were concerned with Standard 2, specifically the integration of diversity activities within lesson planning and teaching practices. The solution has been to find teachers who have succeeded in integrating Standard 2 and use those teachers to model successful integration within lesson planning and teaching practices, which has served as a positive effect on teaching practices and teacher leadership through collaboration.

In multiple interviews, principals were concerned with the lack of an element in the standards that addressed classroom management and student discipline. According to principals, the existing elements in Standards 1 and 4 have been inadequate in addressing classroom management and discipline issues. Principals have coped by using the existing elements to initiate discussions with teachers to improve their classroom management, in turn improving teaching practices. One principal also took issue with Standard 4, specifically the appropriateness of Element H which rated teachers based on their ability to design learning groups among students. This principal was concerned that teachers with at-risk students have found it difficult to achieve higher ratings, which has been detrimental to teaching practices.

Across multiple interviews, principals were concerned about rating teachers within formal classroom observations based only upon what they observed during one particular classroom observation. It was unclear overall how principals were coping with this process, although there was some evidence suggesting principals were considering teacher performance across informal and formal observations on the aggregate when rating teachers during any particular formal classroom observation.

Principals uniformly agreed that Standard 6 and EVAAS® outcomes were being used as a diagnostic to aid in differentiating among teachers based on performance. The

primary use of EVAAS® outcomes was to improve teaching practices at all skill levels. No principal was opposed to the use of EVAAS®; however, the principal at an alternative school was concerned that the use of EVAAS® to rate teachers could be unfair because of the revolving door that exists in students entering and exiting and because of the unique types of students who attend the school.

Principals communicated that they have used formative and summative evaluation of teaching practices and teacher leadership to spur improvement for teachers. However, there were central differences in how principals approach providing feedback and the way in which teachers expected to receive feedback. This resulted in a novel category that was adapted to the existing categories; the novel category was entitled "Principal and Teacher Expectations of Feedback."

Principals expected to initiate discussions that would tap into teachers' past training and professionalism that would lead to improvement in teaching practices and teacher leadership. However, teachers expected specific feedback that related to their pedagogy and subject area. Teachers perceived that principals were unable to provide enough and specific feedback because of a lack of knowledge of their content area.

Teachers were positive overall about the ability of principals to offer specific feedback to improve teacher leadership roles, and principal responses concurred. Across many responses, principals commented that their number one goal was to improve teaching practices and teacher leadership by providing avenues in which teachers could participate in leadership activities.

Chapter 5: Summary, Conclusions, and Recommendations

Summary

The purpose of this study was to explore the effects of NCEES on teaching practices and teacher leadership in six local high schools in a mostly rural southwestern area of the North Carolina piedmont. This study adopted a mixed-methodology (QUAL-quan) process where qualitative data served as the backbone of the study and were collected from teachers through a survey process and from principals through an interview process. Quantitative data analysis was used within the survey process and within a process to examine the ratings of teachers over time and across observed sources.

In Chapter 1, there was evidence provided that reform was necessary in U.S. high schools due to a lack in the quality of teaching practices and teacher leadership. It was also determined that student achievement outcomes were in need of improvement. Alongside the lack in quality of teaching practices, teacher leadership, and student achievement outcomes, teachers were shown to receive high ratings in teacher evaluation systems across the U.S. leading to a contradiction in outcomes.

Chapter 1 also outlined theories relating to a lack of classroom observations for teachers, an "encapsulation" theory reflecting a lack of collaboration among teachers, and a "buffering" in which administrators act as a buffering agent protecting subpar teacher performance.

The following two research questions were chosen.

1. How has NCEES affected teaching practices in six local high schools in the county under study?
2. How has NCEES affected teacher leadership in six local high schools in the

county under study?

In Chapter 2, the literature review, which was delayed until after initial qualitative data were collected through piloted teacher focus groups and principal interviews, refined four categories intially devised from the pilot data using grounded theory while exploring the historic shortcomings of teacher evaluation systems and their effects on teaching practices and teacher leadership. The literature review was used to validate data collection instruments that were used in the primary research methods of this study.

In Chapter 3, the primary research methods of this study were outlined. The primary research included the following steps.

1. NCEES Teacher Survey process.
2. NCEES Principal Interview process.
3. Analysis of NCTSR across years of NCEES application.
4. Analysis of NCTSR across three chosen sources from the literature.

In Chapter 4, the results of the primary research methods were presented. First, the results of the NCEES Teacher Survey were presented in descriptive and inferential terms.

Summary of results–NCEES Teacher Survey process. A table was constructed to give an abbreviated view of the effects of NCEES on teaching practices and teacher leadership as summarized from Chapter 4. Each item from the NCEES Teacher Survey was piloted and validated from an investigation found in the literature review and are displayed in the following table.

Table 14

NCEES Effects Generated from Teacher Unstructured Responses

Structured Items	Effects on Teaching Practices and Teacher Leadership
7. Adequate Formal Observations	• Teachers were overall positive about receiving adequate formal classroom observations. • Teachers viewed formal observations as a way for observers to collect as much data as possible about their classrooms. As a result, teachers responded that they received more feedback and improved their teaching practices and teacher leadership. • No teacher provided a response that they had not received the prescribed number of classroom observations based on their beginning or career status.
9. Adequate Informal Observations	• Teachers were overall positive about receiving adequate informal classroom observations but not as positive when compared to formal classroom observations. • Some teachers doubted whether principals could collect meaningful data during informal classroom observations and questioned the lack of feedback from them. As a result, some teachers responded that informal classroom observations did not affect their teaching practices and teacher leadership. • Although few in number, some teachers responded that they had not received snapshots or walkthroughs.
11. Adequate Time	• Teachers were positive overall about experiencing enough time to engage in pre and postconferences and improved their teaching practices and teacher leadership as a result.
13. Ratings for Standards 1-5 align with 6	• Teachers were slightly negative about the alignment between their ratings for Standards 1-5 and Standard 6 EVAAS® outcomes. Those teachers who experienced misalignment attributed it to the inability of Standard 6 EVAAS® outcomes to properly reflect the learning of their students and the complexity of their classrooms. • Teachers experiencing alignment perceived a validation effect in that they were considered effective teachers across multiple indicators.
15. Appropriateness of Standards 2-5	• Teachers were positive overall about Standards 2-5 on the rubric and provided evidence of internalizing the rubric in order to receive higher ratings from their principal. • By internalizing the rubric and being motivated to receive higher ratings, teachers displayed motivation to improve their teaching practices.
17. Appropriateness of Standard 1	• Teachers were positive overall about Standard 1 on the rubric. • Teachers were concerned about the ability of principals to observe leadership in action because leadership roles were not always accomplished in the classroom. As a result teachers took note of the leadership roles in which they participated in order to relay them to principals during summative evaluations.

(continued)

Structured Items	Effects on Teaching Practices and Teacher Leadership
19. Validity of NCEES–Teaching Practices	• Teachers were positive overall about the ability of Standards 2-5 to measure teacher behaviors that led to student learning. Because of this confidence in the validity, teachers were apt to consider NCEES in planning lessons.
21. Validity of NCEES–Leadership	• Teachers were positive overall about the ability of Standard 1 to measure teacher leadership by design but were still concerned that most leadership roles were not observable.
23. Reliability of Ratings–Standards1-5	• Teachers were positive overall about the reliability of the NCEES rubric. There was little indication that principal ratings were radically different from semester to semester or year to year. As a result, teachers trusted feedback and improved teaching practices and teacher leadership.
25. Reliability of Observers	• Teachers were positive overall about the reliability of principal ratings. As a result, teachers trusted feedback and improved teaching practices and teacher leadership.
27. Validity of EVAAS®–Student Data	• Teachers were negative overall about Standard 6 EVAAS® outcomes reflecting the academic growth of their students. • Teachers with positive perceptions experienced the same validation effect as in structured Item 13 and unstructured Item 14. • Teachers with negative perceptions characterized their students based on a spectrum of abilities of which teachers did not think Standard 6 EVAAS® could appropriately measure growth. • Teachers who viewed Standard 6 EVAAS® outcomes as a reflection of their students' growth experienced validation and carried bright prospects for their ability to increase student learning in the future. • Teaching to the test was a concern shared by teachers across all perceptions.
29. Validity of EVAAS®–Teacher Data	• Teachers were negative overall about Standard 6 EVAAS® outcomes reflecting the quality of their teaching practices. • Teachers with positive perceptions experienced the same validation effect as in structured Item 13 and unstructured Item 14, alongside the same forward-looking perspective from structured Item 27 and unstructured Item 28. • Teachers with negative perceptions viewed their classroom as a complex environment requiring complex teaching practices that Standard 6 EVAAS® could not measure. • Teaching to the test was a concern shared by teachers across all perceptions.
31. Reliability of EVAAS®	• Teachers were slightly negative overall about the reliability of EVAAS® outcomes year to year. • The perception of a lack of reliability in EVAAS® created frustration and confusion among teachers with negative perceptions.

(continued)

Structured Items	Effects on Teaching Practices and Teacher Leadership
33. Unintended Consequences–EVAAS®	• Teachers were slightly positive overall about EVAAS® outcomes influencing them to use teaching practices that helped their students grow academically. • Teachers viewed EVAAS® as a device to add to their already existing repertoire to increase student achievement explaining the contradiction of an EVAAS®-related item receiving a level of positivity when all other were negative. • There were no reports from teachers of a concern that principals would deflate or inflate ratings for Standards 1-5 to make them match EVAAS® outcomes. This unintended consequence appeared in the literature review but not in the primary research outcomes. Observer rating without deflation or inflation provides teachers with valid ways to improve their teaching practices. • Teaching to the test was a concern shared by teachers across all perceptions.
35. Formative Feedback–Practices	• Teachers were negative overall about receiving feedback from observers to improve teaching practices. • Teachers concentrated on expecting specific feedback on ways to improve teaching practices. Because teachers perceived a lack of specific feedback, teaching practices were not affected.
37. Formative Feedback–Leadership	• Teachers were positive overall about receiving feedback from observers to improve teacher leadership. • Teachers received specific feedback about how to improve teacher leadership and did so. • Specific feedback about improving leadership usually centered on improving the functionality of PLCs or the SIT to boost student achievement.
39. Summative Feedback–Practices	• Teachers were negative overall about receiving summative ratings and feedback from observers to improve teaching practices. • Teachers continued with the theme originating in Item 37 that they expected but received a lack of specific feedback relating to an improvement of their teaching practices. • Teachers described time restraints involved in postconferences when summative ratings were provided. The time-restraint resulted in a description of an ‘assembly-line feeling.’
41. Summative Feedback–Leadership	• Teachers were positive overall about receiving summative ratings and feedback from observers to improve teacher leadership. • The same phenomenon was observed as in structured Item 37 and unstructured Item 38. Teachers received specific feedback about how to improve teacher leadership and did so. • Teachers were less positive with summative feedback to improve teacher leadership than formative feedback (Item 37) because of time constraints.

Summary of results–NCEES Principal Interview process. Principals provided

qualitative data that overall paralleled teacher input on unstructured items on the NCEES Teacher Survey. One exception surrounded the use of EVAAS® where principals were attempting to provide solutions to teacher misconceptions. Evidence from teacher responses suggested teachers were not accepting the solutions principals were offering and misconceptions still existed. Another exception existed where the expectation of principals and teachers differed on giving and using feedback during formative and summative feedback. A table was constructed to show an abbreviated list of principal perceptions and how they related across each item of the NCEES Teacher Survey.

Table 15

NCEES Effects Generated From Principal Interviews

NCEES Teacher Survey Item	Effects on Teaching Practices and Teacher Leadership
7. Adequate Formal Observations	• Principals relied on formal classroom observations to collect and compile data for all teachers. It was important to principals that they saw improvement for all teachers, not just beginning teachers.
9. Adequate Informal Observations	• Principals used informal classroom observations for general purposes to check for "I can . . ." statements, lesson planning, and use of technology. • Informal classroom observations served as a probe to check for areas of improvement.
11. Adequate Time	• Principals delegated classroom observation duties to ease time requirements and collect more data. This allowed for more feedback that could help teachers improve teaching practices and teacher leadership. • Although principals agreed NCEES required more time they thought, it was worthwhile because of the focus on improving teaching practices and teacher leadership.
13. Ratings for Standards 1-5 align with 6	• Principals were generally unaware of how their ratings aligned with teachers' Standard 6 EVAAS® outcomes. • Principals used data from Standard 6 EVAAS® outcomes as a diagnostic tool to find out where teachers needed to improve.
15. Appropriateness of Standards 2-5	• Principals agreed that Standards 2-5 contained a spectrum of teacher behaviors that would allow them to offer advice to improve teaching practices and teacher leadership. Principals viewed this as an important way to increase student learning. • Principals provided advice to improve Standards 1 & 4. Principals communicated that Standards 1 & 4 were missing a specific element necessary to properly rate the classroom management skills of teachers. Also, there was concern that not all teachers could fairly be rated in Element F based on allowing students to "create and manage learning teams." (McREL et al., 2012, p. 28)
17. Appropriateness of Standard 1	• Principals agreed that Standard 1 contained important characteristics of teacher leadership they could use to improve teacher leadership.
19. Validity of NCEES–Teaching Practices	• Principals agreed that Standards 2-5 identified important teaching practices that they could use to advise teachers. Principals thought an improvement in teaching practices would produce improved student outcomes.
21. Validity of NCEES–Leadership	• Principals agreed that Standard 1 gave them the ability to improve teacher leadership and collaboration in PLCs to ultimately boost student achievement.

(continued)

NCEES Teacher Survey Item	Effects on Teaching Practices and Teacher Leadership
23. Reliability of Ratings–Standards 1-5	• Principals were generally unaware whether ratings for Standards 1-5 aligned over time.
25. Reliability of Observers	• Principals made a commitment to collect as much data as possible and engaged in discussions with peers about what it meant for teachers to be "Developing," "Proficient," "Accomplished," or "Distinguished." By being accurate raters, principals recognized this would increase accurate feedback and improve teaching practices and teacher leadership.
27. Validity of EVAAS®–Student Data	• Principals credited NCEES with an improvement in teaching practices and teacher leadership to some level. As a result, student outcomes improved. • Principals also cited other variables for improving student outcomes such as recruitment/retention of competent teachers and the collaborative work that was occurring in PLCs.
29. Validity of EVAAS®–Teacher Data	• Principals attributed improvements in teaching practices to using NCEES but without details. • Principals attributed improvements in teacher leadership by using NCEES to identify deficiencies, offering teachers leadership roles, and using PLCs as the vessel in which improvements could proliferate.
31. Reliability of EVAAS®	• Principals responded that they observed some teachers moving from low to high ratings and vice versa with Standard 6 EVAAS® outcomes. • Principals viewed contradicting outcomes as confusing but were confident that valid ratings from Standard 6 EVAAS® outcomes would be produced over 3 years of data.
33. Unintended Consequences–EVAAS®	• Because principals were generally unaware of how their ratings for Standards 1-5 aligned with Standard 6 EVAAS® outcomes, this provided indirect evidence there was no rating deflation or inflation. Direct evidence from the NCTSR reinforced that principals were unbiased by EVAAS® outcomes in rating teacher performance. The lack of inflation or deflation resulted in valid and reliable principal ratings that were trustworthy and could be used to improve teaching practices and teacher leadership.
35. Formative Feedback–Practices	• Principals viewed formative feedback as a way to tap into the experiences of teachers giving teachers the ability to reflect on their own and find ways they could improve their teaching practices. It was important to principals that teachers have the ability to solve problems and generate solutions to potential problems apparent in their classroom. Principals wanted to respect teacher professionalism.
37. Formative Feedback–Leadership	• Principals offered teachers specific ways to improve leadership roles. Improving the ways teachers functioned within PLCs and the SIT were a high priority.

(continued)

NCEES Teacher Survey Item	Effects on Teaching Practices and Teacher Leadership
39. Summative Feedback–Practices	• Principals aimed to combine ratings across formal and informal classroom observations to provide teachers with a final status that would motivate teachers to improve teaching practices. Similar to formative feedback to improve teaching practices, principals wanted teachers to think about issues they were potentially having in their classrooms and generate possible solutions.
41. Summative Feedback–Leadership	• As with formative feedback, principals honed in on specific functions of teachers within PLCs and the SIT to improve teacher leadership.

Summary of results–inferential tests using NCEES Teacher Survey demographic groups. The Mann-Whitney and Kruskal-Wallis tests found significant differences among groups on three items from the NCEES Teacher Survey where the null hypothesis was rejected. On Item 13, teachers with fewer years of experience agreed at a rate higher proportionally that they had enough time to participate in pre and postconferences. On Item 27, teachers with more years of experience and teachers who had experienced a previous teacher evaluation system prior to NCEES (Teacher+1) agreed at a higher rate proportionally that Standard 6 EVAAS® ratings reflected the academic growth of their students. On Item 35, teachers with fewer years of experience and teachers who had not experienced a previous teacher evaluation system prior to NCEES (Teacher+0) agreed at a high rate proportionally that they received useful feedback from observers to improve their teaching practices.

There were no significant differences found among schools or the gender of teachers across items on the NCEES Teacher Survey for which the null hypothesis was accepted. Because of the lack of overall differences in the inferential statistical tests, the quantitative and qualitative analysis using the NCEES Teacher Survey, and responses from principals, this study disconfirmed Lynn et al.'s (2013) findings for the sample of

teachers in the schools under study. While Lynn et al. showed an overrepresentation of White teachers; female teachers; and teachers with more years of experience in receiving higher ratings from principals, there was no indication of any bias among these groups in the primary research of this study.

The pre and posttest design of the demographic groups Teacher+0 and Teacher+1 carried a significant status in the primary research of this study. Because of their status, responses across items on the structured and unstructured items from the NCEES Teacher Survey were analyzed for differences especially on Items 27 and 35 where the Mann-Whitney test uncovered significant differences across the Teacher+0 and Teacher+1 groups causing the null hypothesis to be rejected.

For Item 27, teachers who had experienced a previous teacher evaluation system prior to NCEES (Teacher+1) agreed at a higher rate proportionally that their Standard 6 EVAAS® ratings reflected the academic growth of their students, but there was no trend uncovered when analyzing unstructured Item 28 that explained this difference. Responses from both Teacher+0 and Teacher+1 groups indicated hesitancy in accepting Standard 6 EVAAS® outcomes as valid given the characteristics of students who both groups of teachers described. Teachers across both Teacher+0 and Teacher+1 groups described in a parallel manner the difficulty in accepting Standard 6 EVAAS® outcomes based on the complexity of their classrooms.

The same phenomenon was observed for Item 35 where teachers who had not experienced a previous teacher evaluation system prior to NCEES (Teacher+0) agreed at a higher rate proportionally that they received useful feedback from observers to improve their teaching practices. Responses from both Teacher+0 and Teacher+1 groups indicated a parallel response pattern–both groups desired more specific feedback from principals

about how to improve teaching practices.

One explanation for the failure to detect the differences across Teacher+0 and Teacher+1 groups across Items 27 and 35 may have stemmed from the fact that only a fraction of the teachers who provided responses to the structured items offered responses on unstructured Items 28 and 36 describing the effects on their teaching practices. The motivation of teachers in their responses that caused the significant differences uncovered by the Mann-Whitney test across Items 27 and 35 requires further study in the future.

While analyzing Items 27 and 35 resulted in an unclear relationship between Teacher+0 and Teacher+1 groups in establishing why there were significant differences among groups, it was important to distinguish the insignificant relationship between groups Teacher+0 and Teacher+1 across the structured and unstructured items of the NCEES Teacher Survey. Because of the significant status of the groups Teacher+0 and Teacher +1 in the primary research of this study as the pretest and posttest effects of NCEES on teaching practices and because of the insignificant differences across these two groups on the structured and unstructured items across the NCEES Teacher Survey other than on Items 27 and 35 for which the difference was significant but undetermined as to why, these contradictory results helped determine an overall lack of an effect of NCEES on teaching practices.

Summary of results–Analysis of NCTSR across standards and years. A chi-square analysis of NCTSR across years revealed that principal ratings of teachers have been compressed towards the "proficient" and "accomplished" rating categories with statistical signficant differences across years. Further investigation revealed that ratings for Standards 1-5 were found to have migrated away from "developing" and "accomplished" towards the middle categories. Standard 6 EVAAS® ratings were shown

to be dispersed across years but did show significant differences across years in some cases.

The goal in the analyis of NCTSR was to determine whether there was significant movement in teacher ratings across years to infer whether there was an improvement in teaching practices and teacher leadership. It was notable that because the ratings from principals for Standards 1-5 were compressed into the middle categories did not mean there was a widespread degredation in the quality of teaching practices and teacher leadership. It was determined that principals may have become more accurate raters over time and because of training.

Summary of results–Analysis of NCTSR across three chosen sources. A chi-square analysis showed that the teacher ratings for the schools in the county under study more closely resembled the distribution of ratings from Lynn et al. (2013) reinforcing the observation that teacher ratings from principals have been compressed. A more acceptable distribution of teacher ratings was observed in Curtis (2011), but the teacher ratings in the schools in the county under study significantly differed from the ratings in Curtis (2011).

Conclusions

The effects of NCEES on teaching practices in the six schools in the county under study. From the quantitative and qualitative analysis of the NCEES Teacher Survey, the qualitative analysis of the NCEES Principal Interview process, and the purely quantitative analysis of the NCTSR and distribution of teacher ratings, there was ambiguous evidence that NCEES had a net effect overall on teaching practices in the six high schools in the county under study. The evidence for a net effect balanced out. The following narrative described this conclusion in detail.

Across responses on items from the NCEES Teacher Survey and the NCEES Principal Interview process, the most incriminating piece of evidence driving an ambiguous status for the first research question was that both principals and teachers were unable to produce specific examples of ways teaching practices improved due to NCEES. When principals offered responses to ways in which NCEES affected teaching practices they saw while touring classrooms, principals were unable to identify specific improvements in teaching practices produced by NCEES and instead attributed general improvements to other processes such as the collaboration of teachers within their PLCs and professional development. Principals identified NCEES as a general driver of improvement without providing specific details and stipulated that these improvement activities happened to some degree independent of NCEES.

Teachers responded similarly to principals on unstructured items on the NCEES Teacher Survey. Teachers who responded positively across the items rarely listed a specific teaching practice that was improved due to NCEES but were superficial in their responses that their teaching practices had improved. Improved lesson planning and a validation effect of effective teachers were the most specific examples of how teachers improved their teaching practices as provided on the unstructured items of the NCEES Teacher Survey, but these responses were rare while superficial responses were the rule.

Within the design of NCEES, precaution was taken to ensure the structure of NCEES followed recommendations from recent literature (Danielson, 2007). NCEES has programmed within it the mandate that all teachers receive a certain number of formal and informal classroom observations based on their status as beginning, career, or renewing career teachers (McREL et al., 2012). This approach curtailed the lack of classroom observations that teachers have experienced in the past (NCTQ, 2011a) and

laid the groundwork for a successful teacher evaluation system that historic teacher evaluation systems were missing (Good & Brophy, 1984; McGreal, 1983; Schmoker, 2006). By mandating that all teachers receive classroom observations based on their status as beginning or career status teachers, the stage was set to collect data about teacher performance, distribute teachers based on their performance, and implement changes to improve teaching practices across the spectrum of teacher performance.

Teachers agreed they received adequate formal classroom observations, and not one teacher responded on the NCEES Teacher Survey he/she had not received the prescribed number of classroom observations based on his/her status for the school year under study. However, fewer teachers agreed that they received an adequate number of informal classroom observations possibly meaning observers in the schools under study were prone to Darling-Hammond et al.'s (1983) "insufficient sampling of performance" (p. 306), at least in rating some teachers. A small number of teachers in the sample responded that they received no snapshots or walkthroughs for the school year under study.

Principals concurred with teachers in attempting to provide an adequate number of formal and informal classroom observations. Principals used support personnel to ease time requirements and collect more data. Principals expressed the desire to observe teachers in action as much as possible.

The unstructured items associated with an adequate number of formal and informal classroom observations and adequate time were the first juncture where teachers responded without specificity as to how NCEES affected their teaching practices. Those teachers with positive perceptions left responses that were generic and exemplified by superficial responses such as "helped me to change/improve different aspects of my

teaching"; "I received some pointers"; "giving me direct feedback"; "I strive to be prepared"; and "boosted my confidence." While these responses reflected worthwhile perceptions of teachers that teaching practices improved, they lacked enough supporting details to conclude there was an overall improvement.

The second category under measurement in the NCEES Teacher Survey centered on Standards 1-5 and data collection instruments. Both principals and teachers shared positive overall perceptions of the standards used in data collection instruments and the rubric used to rate teachers. Principals and teachers shared positive perceptions overall of Danielson's (2007) *Framework* as adapted in NCEES to measure a wide range of teaching practices and teacher leadership with the aim of boosting student achievement, which has been a theme suggested throughout the literature (Hanover Research, 2011; Peterson, 2000). Principals and teachers also agreed that the standards and elements in the data collection instruments of NCEES could help differentiate among teachers based on the quality of teaching practices and teacher leadership, which was a second important theme in the literature (Curtis, 2011; Daley & Kim, 2010; Weisberg et al., 2009).

An apparent contradiction arose when analyzing NCTSR across available school years from the online NCDPI database. Between the school years 2011-2012 and 2013-2014, teachers were not distributed across a wide range of performance ratings.

The ramifications of this contradiction were important. The most important goal of revamped teacher evaluation systems under the umbrella of the RttT initiative was to produce a rational distribution of teachers based on their performances across rating categories and among multiple measures of teacher effectiveness (Daley & Kim, 2010; MET, 2010, 2012; TNDOE, 2012). This approach was to counter inflated principal ratings that were customary in historic teacher evaluation systems as noted by Frase and

Streshly (1994) and Weisberg et al. (2009). Without a rational distribution of teachers based on performance and among multiple measures, teacher evaluation systems lacked the ability to offer professional development to teachers who needed it most: Middle performers functioned without the stimulus of growth, and top performers have gone without recognition and their practices without replication (MET, 2010, 2012).

While principals and teachers were positive overall across items on the NCEES Teacher Survey about the appropriateness of Standards 1-5, the validity and reliability of Standards 1-5, and the reliability of principal ratings, the compression of principal ratings contradicted their perceptions. If Standards 1-5 were appropriate, valid, reliable, and reliably used by principals, principal ratings for teachers should be distributed along a wide spectrum of teacher ability as in Curtis (2011) and not tightly compressed in the middle as was the case for the principal ratings for the teachers in the schools under study.

The distribution of principal ratings of teachers in the county under study differed at statistically significant levels (α=0.001) from the distribution presented in Weisberg et al. (2009) and Curtis (2011; see Table 13). Upon further analysis, the distribution of principal ratings of teachers in the county under study differed based on compression in the middle rather than at the top as with Weisberg et al.

Also, although the distribution of principal ratings for teachers in the county under study was significantly different than those found in Curtis (2011), the χ^2 statistic displayed drastically lower differences (see Table 13). The distribution of principal ratings for teachers in the county under study versus Weisberg et al. (2009) and Curtis (2011) is displayed in Figure 11 below.

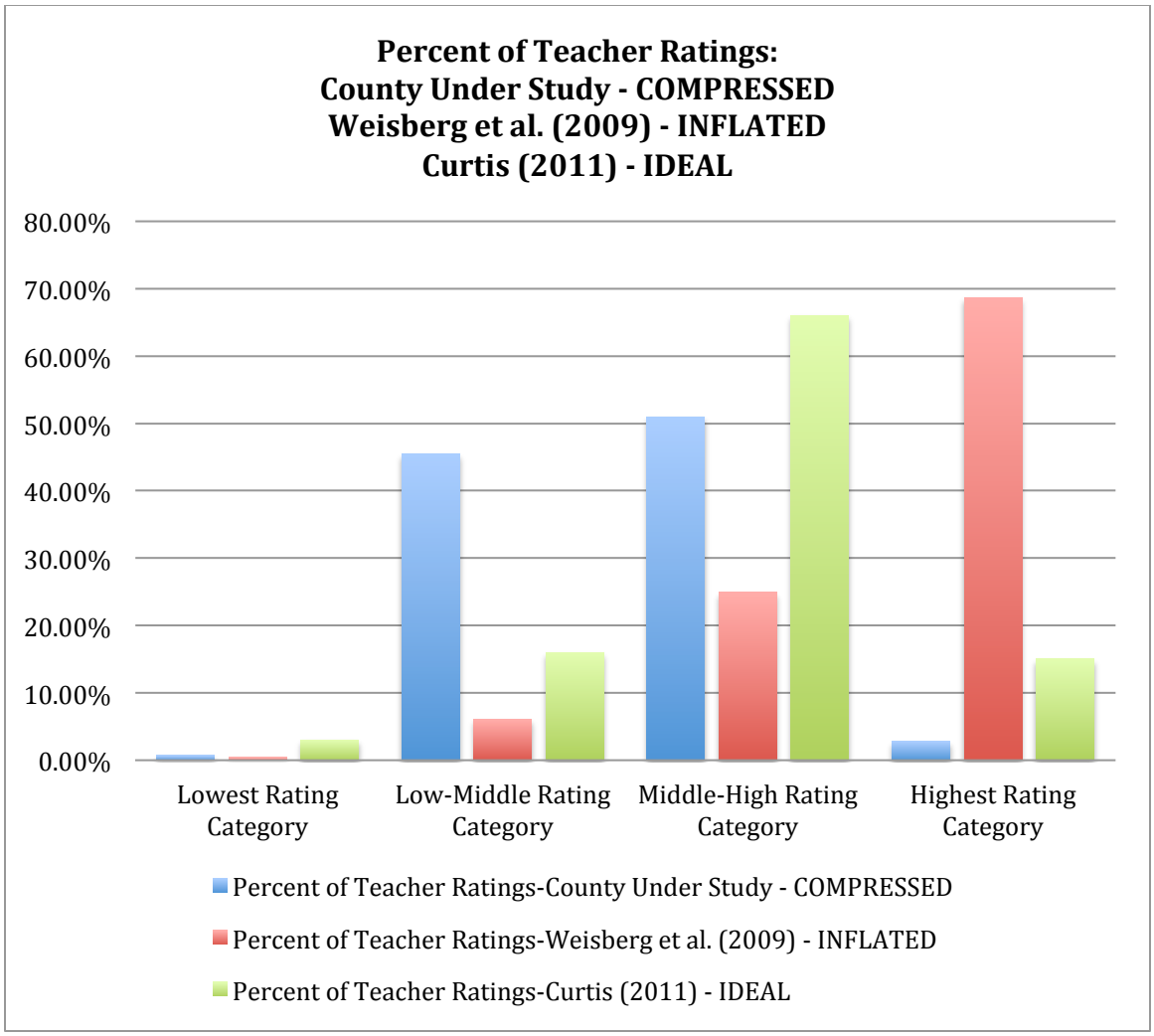


Figure 11. Distribution of Teacher Ratings–County under Study versus Two Sources.

Compounding the problem with the compression of principal ratings for teachers in the county under study were the responses of teachers to unstructured items on the topics of the appropriateness, validity, and the reliability of Standards 1-5, alongside the reliability of observers rating teachers based on Standards 1-5. As with the previous unstructured responses, teachers and principals failed to mention specific teaching practices that were improved and responses were general without specificity. Such comments included implementing different types of instruction, a desire to improve

teaching practices, setting clear goals, finding areas where improvement was needed, improving lesson planning, paying more attention to the standards, confirming what works well, and prioritizing changes. While all of these responses portrayed worthwhile outcomes, they lacked specificity; and when compared to the contradiction apparent in the compression of principal ratings for teachers in the county under study, a decision to find an overall improvement was clouded.

The use of EVAAS® outcomes was the third category under measurement within the NCEES Teacher Survey and presented the only category of the NCEES Teacher Survey in which teachers were consistently negative overall, although most EVAAS®-related items were slightly negative. Teacher perceptions of Standard 6 EVAAS® on unstructured items were as divided as the structured items which diminished the researcher's ability to determine if there was an overall improvement in teaching practices stemming from EVAAS® outcomes. Teaching to the test was a phenomenon that teachers regularly cited within a negative context, and the literature regularly referred to the phrase pejoratively (Collins, 2014; Darling-Hammond et al., 2011; Jones & Egley, 2007).

Similar to the ability of standards-based teacher evaluation systems, VAM has been instituted as a measure to distribute teachers based on their performance across a spectrum of rating categories (TNDOE, 2012). Following an approach from the TNDOE (2012) study, the results of the distribution of EVAAS® ratings for the teachers under study were compared to the distribution of principal ratings for teachers using Standards 1-5 within NCEES. The analysis used the last year of released EVAAS® data released in conjunction with principal ratings, which was for the 2013-2014 school year. The analysis excluded one school that did not have enough teachers rated under Standard 6 to

post results in the NCDPI database. The following graph displays the differences in distribution from EVAAS® ratings for teachers compared to principal ratings using Standards 1-5 from NCEES for teachers in the county under study, with the middle ratings of "proficient" and "accomplished" combined because of their equivalence to the rating "meets expected growth" from EVAAS®.

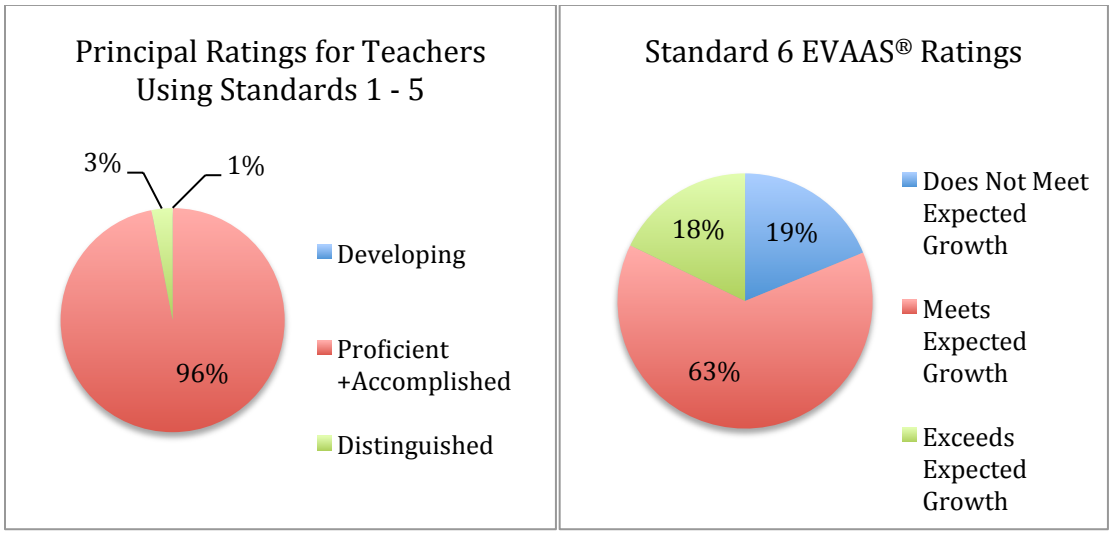


Figure 12. Principal and EVAAS® Ratings for Teachers in County under Study.

Although EVAAS® outcomes distributed teacher ratings across a more rational spectrum of teacher performance, EVAAS® was designed purposely to designate a proportion of teachers as low and high performers by comparing teachers to each other and considering teachers at the bottom and top percentile levels of performance (Glazerman et al., 2010). Proponents have advocated such a competitive approach among teachers to produce higher learning gains on the basis that competition among teachers would spur improvement in teaching practices (Glazerman et al., 2010; Goe, 2008). The important idea was that EVAAS® has been programmed to distinguish among teachers across a spectrum of performance by default, whereas no such programming could be

accomplished with principals in their intuitive judgments of teacher performance.

The competitive approach among teachers to improve teaching practices and produce learning gains was somewhat evident in Item 33 on the NCEES Teacher Survey where teachers responded slightly positively overall that Standard 6 EVAAS® influenced teachers to seek and use practices that would help their students grow academically. However, as with responses on other items throughout the NCEES Teacher Survey, in unstructured Item 34, teachers failed to include specific details as to how their teaching practices improved because of using EVAAS®. Teachers with positive perceptions of EVAAS® outcomes perceived validation by being considered effective teachers across multiple measures, but the perception of validation was vague overall without details of how teachers improved their teaching practices.

Although Standard 6 EVAAS® ratings distributed teachers across a more rational spectrum of teacher performance and teachers with positive perceptions involving EVAAS®-related items perceived a validation effect, these positive outcomes were offset by principal ratings that were not aligned with EVAAS® ratings, slightly negative perceptions of teachers overall about EVAAS®-related topics, and the concern of teachers about teaching to the test used in a negative tone. Most importantly, teachers with positive perceptions in structured and unstructured items on the NCEES Teacher Survey did not offer specific ways in which their teaching practices improved because of Standard 6 EVAAS®. These factors seemed to offset and create an ambiguous conclusion as to whether EVAAS®-related outcomes caused an improvement in teaching practices for the teachers in the county under study.

The last category under analysis in the NCEES survey was, “How Are Outcomes from Teacher Evaluation Systems Utilized?” Evidence was presented from the structured

and unstructured items on the NCEES Teacher Survey and the Principal Interview process that indicated the outcomes from NCEES were not utilized properly in the county under study. A feedback loop between principals and teachers went awry due to a difference in expectations of principals and teachers in giving and receiving feedback. Principals wanted to act as facilitators of learning in the classroom without intruding on the professional decision-making and problem-solving skills of teachers, whereas teachers expected specific feedback to improve teaching practices.

Formative evaluation has been supported and advocated in the literature as a powerful agent of change in teaching practices (Darling-Hammond & Adamson, 2013; Donaldson & Peske, 2010; Sartain et al., 2011). Through targeted, "in-the-moment" (Donaldson & Peske, 2010, p.11) feedback, improvements in teaching practices have been supported by evidence of increased student learning (Kane et al., 2011; MET, 2012). Teachers with positive perceptions of formative evaluation responded that the feedback they received helped improve their teaching practices but without detail.

Teacher responses to the structured and unstructured items from the NCEES Teacher Survey indicated that formative evaluation was not a pathway to improving teaching practices, and it was determined by analyzing teacher responses that the failure of formative evaluation to succeed was because of the observed differences in expectations that both parties held. The result of this breakdown was theorized to extend throughout the processes within NCEES to improve teaching practices–teachers holding positive and negative perceptions about the usefulness of the feedback they received responded to unstructured items across the NCEES Teacher Survey with a lack of specificity in attempting to describe the effects of NCEES on their teaching practices. It appeared as though the barrier concerning the expectations of feedback between

principals and teachers in this category was manifested throughout unstructured items on the NCEES Teacher Survey.

Summative evaluation within NCEES was looked upon less favorably than formative evaluation with a higher rate of disapproval. It was determined by an analysis of responses that the short duration of postconferences led teachers to receive less useable feedback from within the summative framework of evaluation.

Because of the difference in expectations of feedback by principals, the lack of specific effects of the usefulness of feedback on teaching practices, and the negative perceptions of teachers overall on unstructured and structured items associated with how outcomes were utilized, this category caused negativity overall and also caused concern. The concern stemmed from the theory that the difference in expectation of how feedback was given and received between principals and teachers extended throughout responses from teachers on unstructured items across the NCEES Teacher Survey.

In summary, there was ambiguous evidence of a net effect of NCEES on teaching practices. The negative effects seemed to balance the positive effects for an overall neutral result.

The best examples of the positive effects of NCEES on teaching practices were the positive perceptions shared overall by teachers across items measuring the first two categories in the NCEES Teacher Survey–specifically the quantity of classroom observation and teacher evaluation and the quality of classroom observation and teacher evaluation. Although teachers provided generalizations about how their teaching practices had improved across the first two categories measured on the NCEES Teacher Survey, there was a lack of specificity possibly stemming from a breakdown between the expectations of principals and teachers in giving and receiving feedback.

Categories three and four measured on the NCEES Teacher Survey were perceived as slightly negative overall and counteracted the positive perceptions of teachers for the first two categories. Also, contradictions stifled the potential of NCEES in causing positive effects overall, such as the compression of principal ratings for teachers and the misalignment of principal ratings for teachers using Standards 1-5 and Standard 6 EVAAS® ratings.

The anatomy of NCEES. An anatomical analogy could be used to describe the effects of NCEES on teaching practices. To have a healthy person requires biological systems working in concert to maintain homeostasis. To have a healthy teacher evaluation system requires the validated categories generated within the pilot data and literature review working in concert: There should be an adequate amount of classroom observations and available time for pre and postconferences; data collection instruments should be appropriate, valid, reliable, and observers using them should do so reliably; VAM measurements should be valid, reliable, and not cause unintended consequences; and outcomes should be used to spur teacher improvement through formative and summative feedback.

The evidence from the primary research suggests that the structural components of NCEES are healthy. The structural components of teacher evaluation systems are analogous to the skeletal and muscular systems of a human body that provide structure, support, and movement to a healthy person. NCEES has structure, support, and movement in providing adequate formal and informal classroom observations while providing enough time for principals and teachers to operate.

The NCEES standards and data collection instruments are analogous to healthy circulatory and respiratory systems where the standards and instruments are the heart of

NCEES and help exchange information among stakeholders. Evidence from the primary research suggests that the heart and associated processes of the circulation of blood and exchange of gases are occurring within healthy limits. According to teachers and principals, Standards 1-5 of NCEES are appropriate, valid, reliable, and used by observers reliably allowing for the pumping and circulation of data throughout NCEES.

Standard 6 EVAAS® outcomes are analogous to the lymphatic and endocrine systems in that they serve as a diagnostic feature that functions overall to regulate and protect processes. This analogous function of EVAAS® outcomes in biological systems was parallel to its function in the literature to provide another measure of teacher effectiveness (MET, 2012) to help gauge the validity and reliability of other measures such as the principal ratings for teachers (Kane et al., 2010).

EVAAS® outcomes for the schools in the county under study have served as a diagnostic tool for teachers who shared positive perceptions on EVAAS®-related items from the NCEES Teacher Survey. Principals reinforced the use of EVAAS® outcomes as a diagnostic, and two independent sources of data supported the lack of deflation or inflation of teacher ratings based on pressures from EVAAS® outcomes.

However, EVAAS®-related items on the NCEES Teacher Survey were slightly negative overall; and while the use of VAM and EVAAS® has been debated in the literature, teachers with negative perceptions persistently embraced misconceptions surrounding EVAAS®-related topics and issues despite the training they received. This indicates issues within this category for NCEES in the county under study and for the analogous biological lymphatic and endocrine systems. The analogous lymphatic and endocrine systems are not in disrepair; they require further diagnosis of symptoms and treatment.

The outcomes of teacher evaluation systems and their utilization form the final analogy within the comparison of teacher evaluation systems and biological systems, and it is in this category and analogous biological system that presents the largest area of concern. The largest piece of evidence in a determination against NCEES contributing to an overall improvement in teaching practices was the inability of teachers and principals to provide specific teaching practices that improved. Because of the different expectations of principals and teachers in giving and receiving feedback, the effects of the feedback offered by principals were not reaching teachers and improving teaching practices.

This phenomenon is analogous to an inoperable nervous system with synaptic issues in the brain locale where principals and teachers form between synapses analogous to neurons. At one neuron, principals are providing feedback and an analogous neurotransmitter that is not compatible with another neuron formed by teachers. The analogous synaptic bridge is broken resulting in a feedback loop that is unable to transfer data to the rest of the body. As a result, there is a loss of control in other parts of the body and analogous teacher evaluation system to improve teaching practices, which was manifested by subjective responses about how teaching practices had improved throughout the unstructured responses across items on the NCEES Teacher Survey and NCEES Principal Interview process.

The general responses from teachers and principals about improvements to teaching practices provided evidence that the rest of the processes involved in NCEES are working independent of the feedback loop. While there was a lack of definitive improvement in teaching practices brought about by NCEES, principals and teachers were able to identify specific ways in which teachers received feedback from principals;

processed the feedback; and made changes to improve their teacher leadership roles. While the synaptic bridge appears to be broken for the teachers in the county under study in improving their teaching practices, the feedback loop between principals and teachers was intact with working neurons, neurotransmitters, and associated synapses to improve teacher leadership.

The effects of NCEES on teacher leadership in the six schools in the county under study. Unlike the ambiguous results of NCEES in affecting teaching practices, NCEES had a clear and positive effect on teacher leadership. Results from structured and unstructured items on the NCEES Teacher Survey and from principals in the NCEES Principal Interview process revealed a conclusion that sharply contrasted the ambiguous effects of NCEES on teaching practices.

Although analyzing unstructured responses (Items 28 and 36) across Items 27 and 35 resulted in the inability to distinguish why the Teacher+0 and Teacher+1 groups were found to be significantly different using the Mann-Whitney test, the results across all items on the NCEES Teacher Survey for pre and posttest groups indicated teachers experiencing previous teacher evaluation systems perceived a difference in their motivation to improve their teacher leadership and were satisfied with the results. Pretest group Teacher+0 perceived the importance in teacher leadership to collaborate with other teachers to improve student outcomes schoolwide, and posttest group Teacher+1 indicated they were not only more motivated to pursue leadership opportunities than under previous teacher evaluation systems, they indicated satisfaction in receiving credit under NCEES for their leadership roles which went unnoticed in previous teacher evaluation systems and realized the importance of teacher leadership in increasing student achievement schoolwide.

Teachers and principals concurred with the general definition used within this study that teacher leadership mostly involved what teachers did outside their classroom to improve their peers and increase school-wide student learning (McREL et al., 2012; TLEC, 2011). However, there were some responses from teachers who indicated the complexity of teacher leadership and that teacher leadership had not been well defined by the NCEES rubric. The argument was that in many cases, teacher leaders carried out activities that were not observable, although teachers shied away from listing specific activities to which they referred.

The largest determining factor in the judgment that NCEES clearly and positively affected teacher leadership was the specificity in which teachers with positive and negative perceptions responded to unstructured items on the NCEES Teacher Survey. Evidence of the clear and positive effects of NCEES on teacher leadership was primarily revealed by analyzing the discrepancy between structured items measuring the negative perception of teachers receiving feedback to improve teaching practices compared (Item 35) to positive perceptions of teachers receiving feedback aimed at improving teacher leadership (Item 37). An analysis of unstructured items provided evidence that teachers who were initially negative about their perceptions in receiving useful feedback to improve their teaching practices changed their perception in receiving useful feedback to improve their teacher leadership.

This same phenomenon was observed across Items 39 and 41 where teachers changed their negative perceptions about receiving useful feedback from summative evaluation to improve their teaching practices. Teachers who were initially negative in receiving feedback to improve teaching practices (Item 39) changed their perception in Item 41 because of the specific feedback they received from principals to improve their

teacher leadership. However, teachers agreed less that they received feedback to improve their teacher leadership during summative evaluation (Item 41) compared to formative evaluation (Item 39) because of time restrictions apparent during postconferences.

The changing perceptions of teachers across items measuring useful feedback to improve teaching practices compared to teacher leadership provided the clearest evidence that NCEES had cultivated an environment where teachers had improved their teacher leadership roles in the schools in which they taught and beyond at the district and state levels. Teachers with positive perceptions across items were clear in that they were informed of specific leadership opportunities by their principals, and teachers revealed they were motivated in pursuing those opportunities to improve their ratings. Teachers expressed that their goal was to ultimately improve student learning, but responses from teachers involved a perception that their involvement in leadership activities would also increase their ratings for Standard 1 while improving student learning.

Teachers perceived a level of specificity in direction from principals to improve their professional leadership in their PLCs and school improvement activities. Teachers responded on unstructured items associated with measuring leadership that they were pleased overall to pursue these goals and that peer teachers would improve and student achievement would improve as a result. Teachers perceived an increased direction from principals to become actively engaged in the mentoring process of beginning teachers, accepting student teachers into their classroom to mentor them, and a general call to accept responsibilities beyond normal expectations that principals allotted.

The idea that student achievement would improve under improved teacher leadership was supported in the literature (Louis, Leithwood, Wahlstrom & Anderson, 2010; Nappi, 2014). Louis et al. (2010) found that a PLC engaged in collaborative work

significantly and positively affected student achievement in mathematics.

The approach of principals in advising teachers to improve teaching practices and teacher leadership of their peers through PLCs and SIT was a common theme throughout the literature (Blase & Blase, 2000; Brezicha, Bergmark, & Mitra, 2014; Spillane, 2006; TLEC, 2011). Brezicha et al. (2014) defined principals using teachers to influence their peers as "distributed leadership" (p. 99) and focused on developing the ability of teachers to cultivate improvement in themselves and their peers through collaborative processes. TLEC (2011) also concentrated on teacher leaders as stakeholders who were committed to improving their peers through collaborative efforts.

Teacher responses on unstructured items from NCEES displayed their motivation for pursuing roles that placed them front and center in the process of developing their peers. Most importantly, after tracking responses from teachers based on their years of experience and whether they had received classroom observations and evaluation using a prior teacher evaluation system to NCEES, more experienced teachers and teachers who had experienced a previous teacher evaluation system shared a sense of relief and satisfaction that their leadership roles were finally being recognized within NCEES.

Teachers who responded with negative perceptions of experiencing adequate classroom observations and time to participate in pre and postconferences; the appropriateness, validity, and reliability of Standards 1-5; and receiving feedback to improve teacher leadership, offered valuable input on unstructured items across the NCEES Teacher Survey. Although these teachers shared negative perceptions of improving their leadership across categories on the structured items for the NCEES Teacher Survey, they had specific knowledge of leadership opportunities but felt shunned in either not being chosen for these leadership positions or not fully understanding how to

pursue them. The responses from these teachers with negative perceptions about the feedback they received to improve their leadership roles contributed to the theory that teachers were receiving specific feedback to improve their leadership in contrast to teachers with negative perceptions of improving teaching practices under NCEES where teachers responded that they received a lack of specific feedback.

Teachers with positive perceptions of the effects of NCEES to improve their teaching practices were also interested in improving their leadership roles, and these teachers internalized the NCEES rubric. Teachers paid specific attention to advice from their principals, to Standard 1 from the NCEES rubric, and how both used in combination could improve their ratings and improve student learning.

In summary, in analyzing the changing perceptions of teachers across items measuring their ability to improve their teacher leadership compared to their teaching practices offered the most valuable evidence of an improvement in teacher leadership due to the influences of NCEES. Also, teachers with positive and negative perceptions across items provided evidence that teachers were offered specific feedback on how to improve their leadership, and they were motivated to do so, although some teachers felt shunned by not being elected or not knowing how to fulfill the advice of principals to improve their teacher leadership.

Validity, Reliability, Generalizability, and Limitations of Data Collection Instruments

The NCEES Teacher Survey and NCEES Principal Interview items were initially checked for face validity (Gay et al., 2009) prior to a pilot program. The pilot program consisted of two phases with the first involving two teacher focus groups and two principal interviews using teachers and principals from outside the sample used within

the primary study. The first phase of the pilot used open-ended items for teacher focus groups and principal interviews and operated to offer initial data used to construct the NCEES Teacher Survey and NCEES Principal Interview. The initial data gathered from the first phase of the pilot also served to guide the literature review, an approach borrowed from Charmaz (2006).

A second phase of the pilot program analyzed the NCEES Teacher Survey and NCEES Principal Interview for content, sampling, and item validity (Gay et al., 2009) and involved a panel of five teachers and one principal. The second phase of the pilot occurred before the primary research and formal distribution of the NCEES Teacher Survey and the NCEES Principal Interview process.

Although teacher focus groups were used in the first phase of the pilot, they were not repeated in the primary research. Unstructured response items were used on the NCEES Teacher Survey during the primary research to elicit qualitative data from teachers rather than focus groups because of logistical considerations.

The literature review was used to refine categories initially generated from the first phase of the pilot program and provided the basis for the four categories and connected conceptual questions measured by the NCEES Teacher Survey and NCEES Principal Interview instruments used within the primary research methods. The piloting and use of a delayed literature review followed the principles set forth in Charmaz (2006) and Gay et al. (2009).

The construct validity of the NCEES Teacher Survey and NCEES Principal Interview items was built using the literature review, while the second phase of the pilot program provided content, sampling, and item validity (Gay et al., 2009). With the exception of one novel category and four conceptual questions, the responses from

teachers on unstructured items on the NCEES Teacher Survey and principals from the NCEES Principal Interview process were fully coded for and encompassed within the original four categories. This outcome provided evidence of theoretical saturation (Charmaz, 2006) of the original categories but also called for further investigation in the future relating to the novel category "Principal and Teacher Expectations for Feedback." Four novel conceptual questions also called for future investigation. Although an attempt was made to analyze the novel category and four novel conceptual questions with the given data, the NCEES Teacher Survey and NCEES Principal Interview process were not designed to measure them. As a result, future studies need to be undertaken to understand the effects of the novel category and novel conceptual questions within teacher evaluation systems.

While validity was central, reliability was also of importance to support the validity of data collection instruments involved in the primary study (Cohen et al., 2007). Cronbach's Alpha and the split-half coefficient for reliability were calculated, the first using SPSS and the latter an EXCEL spreadsheet. Responses were removed for teachers who had not received Standard 6 EVAAS® ratings (Items 23-33) or had not received enough data for Standard 6 EVAAS® ratings. Also, responses were removed for teachers who had not received enough data for Standards 1-5 (Item 13) to form a perception.

Both alpha and split-half measures were calculated to provide multiple measures of reliability, although split-half measures have been found less accurate (Korb, n.d.). The calculated Cronbach Alpha was 0.87 and split-half coefficient (using the Spearman Brown formula) was 0.84, respectively. Both measures fell within Cohen et al.'s (2007), "highly reliable" category found between 0.80 and 0.90.

Measuring the validity and reliability has justified the basis of data collection

instruments for studies in the past including the present study; but generalizability has presented issues in survey studies, specifically sample sizes (Cohen et al., 2007). The percent of the population (N=299) who responded to the NCEES Teacher Survey presented the largest limitation of the primary research.

The sample (N=90) of responding teachers was 30.1%, falling well short of the required response rate advised by Cohen et al. (2007) which was 48% for a population of N=300. With a sample of N=168 and response rate of 48% for a population of N=300, a researcher could be 95% confident within a confidence interval of +/-5 that the results of a survey were valid (Cohen et al., 2007). This study failed to reach that minimum; and the generalizability of the results within this study to the county under study, other counties in the state of North Carolina, or other counties or states using teacher evaluation systems similar to NCEES should be cautioned although not entirely abandoned.

According to Gay et al. (2009) and Cohen et al. (2007), it has not been uncommon for postal and internet-based surveys to receive low response rates, specifically in the range of 10% to 50% for internet-based surveys. While the response rate for the primary research of this study was considerably higher than 10% at 30.1%, the response rate relegated inferences across sampled teachers rather than upwards at the population level of teachers in the county under study.

Other important limitations were the number of responding teachers who lacked years of experience and the low number of Black teachers responding. The low response rate of teachers with 20 or more years of experience and Black teachers dampened the ability to use inferential statistics to study differences among demographic groups.

Implications and Recommendations

Based on a response rate of 30.1% compared to an advised rate of 48% from Cohen et al. (2007), the implications and recommendations of this study should be understood in the correct context. Although the response rate of the NCEES Teacher Survey does not allow for a desired 95% confidence level coupled with a low confidence interval, the results were robustly detailed for the provided sample. Also, the qualitative responses from teachers and principals could provide valuable insight on the effects of NCEES on teaching practices and teacher leadership regardless of the statistical level of generalizability to the county under study or any other entity.

The implications of this study surround the effects of revamped teacher evaluation systems under the RttT initiative. While traditional teacher evaluation systems were shown in the literature review to lack in the quantity of classroom observations and evaluation, in the quality of classroom observations and evaluation, in the use of multiple measures of teacher effectiveness, and in using outcomes to improve teaching practices and teacher leadership, this study showed that the aim of revamped teacher evaluation systems to improve upon these processes has been realized within NCEES. However, just because the requirements of the RttT initiative were met within its processes does not mean NCEES has been successful in improving teaching practices and teacher leadership.

The recommendations from this study consider shortcomings in the ultimate goal of NCEES to improve teaching practices and teacher leadership. The first recommendation would be for counties and districts in states that have adopted teacher evaluation systems under the RttT initiative to continuously monitor the time demands inherent in adequately observing and evaluating teachers to improve teaching practices and teacher leadership.

Although this study uncovered principals and teachers operating within

acceptable time demands, principals leveraged resources to efficiently use NCEES. Some of those resources were temporary with no guarantee of continuance. The resources principals used in the county under study may not be accessible in other counties across North Carolina or other counties, districts, and states that have adopted an approach to teacher evaluation systems based on the RttT initiative.

Given the evidence found in the pilot data and literature review, a lack of time to carry out classroom observations and pre and postconferences has the ability to seriously damage the ability of teacher evaluation systems to positively affect teaching practices and teacher leadership. Principals in the county under study were innovative in how they approached this issue, but this study was cross-sectional over 1 school year. As a result, continually monitoring the time demands of principals and teachers should remain an important goal.

A second recommendation focuses on opening conversation about the perception of principals in this study that Standards 1-5 of the NCEES rubric lack a dedicated element to rate teachers based on classroom management. Principals are using existing elements (primarily Standard 1 Element A) to begin conversations with teachers about classroom management, but principals communicated the existing standards and elements may be inadequate to begin fully functional conversations with teachers about how they manage classrooms and diffuse behavioral issues. Principals who expressed this perception were concerned about NCEES negatively affecting teaching practices and teacher leadership by not fully communicating expectations of teachers in managing classrooms and potential discipline issues, especially for beginning teachers.

A third recommendation is also based on principal perceptions of Standards 1-5 to improve teaching practices and teacher leadership. Many principals were concerned with

the guideline that they use the NCEES rubric to rate only what they observe in the classroom during the time of a classroom observation.

The principals who expressed concern with this guideline were finding it difficult to observe all the elements necessary to rate a teacher during one classroom observation and finding it difficult to begin conversations about improving teaching practices and teacher leadership. Either professional development of principals or a change in an approach is necessary to address this perception communicated by multiple principals within this study.

A fourth recommendation is to professionally develop teachers more specifically about the diagnostic strengths of EVAAS® data in order to improve teaching practices. Most EVAAS®-related items on NCEES received negative responses in the quantitative and qualitative analyses. Few teachers responded on the unstructured items that they were tapping into the potential of using EVAAS® outcomes to gauge their effectiveness with low, middle, and high student achievers. Compounding the lack of using EVAAS® as a diagnostic, teachers were holding misconceptions of EVAAS® outcomes being influenced by student characteristics and EVAAS® estimates of teacher effectiveness being based on students taking one test on 1 day of the semester or year.

To improve teaching practices, a fifth recommendation would be for the state and counties in North Carolina to consider the approach and effects of using EVAAS® outcomes in determining a final status for teachers. According to NCDPI, a 3-year rolling average of EVAAS® outcomes will be used during the 2015-2016 school year to determine a final status for teachers on Standard 6 based on 3 years of EVAAS® data (NCDPI, 2013). If a teacher's 3-year rolling average results in a rating of "in need of improvement," the teacher could be rated the same on an overall status regardless of the

teacher's ratings on Standards 1-5 (NCDPI, 2013).

Given the evidence from teachers on unstructured items from the NCEES Teacher Survey involving EVAAS®, teachers are concerned about the focus on teaching to the test and the ability of EVAAS® to measure teacher effectiveness given the complexity of the classroom. To deter the approach of teaching to the test and to focus on using EVAAS® as a diagnostic, it would be advisable for the state of North Carolina to follow MDCPS in making EVAAS® outcomes a percentage of a teacher's yearly rating and final status (MDCPS, 2015). MDCPS uses EVAAS® outcomes as 35% of a teacher's yearly performance evaluation without using EVAAS® as a possible sole determinant of a teacher's performance evaluation. Using EVAAS® outcomes as a possible sole determinant of a teacher's performance evaluation contradicts sources cited in the literature review of this study (ASA, 2014; Braun, 2005; NAP, 2009).

The final recommendation centers on communication lines within the feedback loop between principals and teachers. In order to realize NCEES as an agent of change in improving teaching practices, attention needs to be given to the coaching conversations between principals and teachers.

This study uncovered a barrier that exists in the mindset of teachers in the expectations they hold in receiving feedback from principals. Principals in the county under study revealed evidence that they were following Danielson's (2010/2011) approach in coaching teachers to reflect and think about what they could do to improve teaching practices.

Rather than accepting feedback from principals in a manner that stimulates self-reflection and problem solving, teachers expected specific feedback from principals to improve teaching practices. This is a pivotal, final step in improving teaching practices.

Evidence uncovered in this study found that professional development is necessary for teachers to realize the goal of NCEES in improving teaching practices. Professional development for teachers should center on using reflection from principal feedback to improve teaching practices.

Recommendations for Further Study

Several areas for future studies were revealed within the primary research of this study. One future study could center on collecting evidence from principals about how Standards 1-5 of NCEES has led to improvements in teaching practices and teacher leadership. The study could also collect data about how to improve the standards if necessary.

Another future study could center on the effects of using EVAAS® to estimate teacher effectiveness on teaching practices and teacher leadership. Such a study could collect qualitative perceptions from teachers to investigate how teachers are responding to EVAAS® outcomes in their planning, instructional processes, assessments, and what supports teachers are utilizing to improve EVAAS® outcomes.

Finally, a future study could investigate the breakdown in the feedback loop between principals and teachers in using NCEES as revealed in this study. The study could collect qualitative perceptions from principals and teachers about what is expected during the feedback process. Although the primary research methods of this study uncovered a barrier in the feedback loop, neither the NCEES Teacher Survey or NCEES Principal Interview process were fully designed to measure the breakdown in the feedback loop. As a result, future investigation is necessary in this area.

Summary

NCEES was found to be designed with functional processes to overcome the

ineffective nature of traditional teacher evaluation systems uncovered in the literature review. Principals were involved in classroom observations to a higher extent than found in traditional teacher evaluation systems within the literature review. With principals visiting classrooms more often, an important foundation has been built to collect ample data and provide feedback to improve teaching practices and teacher leadership.

Principals and teachers expressed an overall acceptance of Standards 1-5 to rate teachers and provide teachers feedback to improve teaching practices and teacher leadership. The use of EVAAS® outcomes fulfilled the call for revamped teacher evaluation systems to use multiple measures of teacher effectiveness to improve teaching practices. Although teachers held an overall negative perception of EVAAS®-related items on the NCEES Teacher Survey, their negative perceptions were not overwhelming.

Because of a breakdown within the feedback loop between principals and teachers, teachers were not using feedback from principals to improve teaching practices. The breakdown in the feedback loop between principals and teachers should be the subject of further studies.

NCEES was found to improve teacher leadership for the teachers in the county under study. The improvement stemmed from formative and summative feedback that teachers accepted and utilized to improve their leadership roles in PLCs and school improvement activities. However, because of different expectations held by principals and teachers and because of balancing positive and negative effects overall, NCEES was found to have no net effect on teaching practices for the teachers in the county under study.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago Public High Schools. *Journal of Labor Economics, 25*(1), 95-135. Retrieved from http://faculty.smu.edu/millimet/classes/eco7321/papers/aaronson%20et%20al.pdf

American College Testing. (2009). *ACT national curriculum survey 2009* (Rep.). Retrieved from http://www.act.org/research/policymakers/pdf/NationalCurriculumSurvey2009.pdf

American College Testing. (2011). *The condition of college and career readiness* (Rep.). Retrieved from https://www.act.org/research/policymakers/cccr11/pdf/ConditionofCollegeandCareerReadiness2011.pdf

American Institutes for Research & American Federation of Teachers. (2011, April 8). *Workplaces that support high-performing teaching and learning* (Rep.). Retrieved from http://www.air.org/sites/default/files/downloads/report/GenerationY_Workplaces_That_Support_Teaching_and_Learning_0.pdf

American Statistical Association. (2014, April 8). *ASA statement on using value-added models for educational assessment* (Rep.). Retrieved from http://www.amstat.org/policy/pdfs/asa_vam_statement.pdf

Amrein-Beardsley, A., & Collins, C. (2012). The SAS Education Value-Added Assessment System (SAS® EVAAS®) in the Houston Independent School District (HISD): Intended and unintended consequences. *Education Policy Analysis Archives, 20*(12). Retrieved from http://epaa.asu.edu/ojs/article/view/1096

Anderson, H. M. (1954). A study of certain criteria of teaching effectiveness. *The Journal of Experimental Education, 23*(1), 41-71. Retrieved from http://www.jstor.org/stable/20153950?seq=1#page_scan_tab_contents

Appeldoorn, K. L. (2005). *Developing and validating the collaboratives for excellence in teacher preparation (CETP) core evaluation classroom observation protocol (C.O.P.)* (Doctoral dissertation, University of Minnesota, 2004) (pp. 1-219). Ann Arbor, MI: Proquest Information and Learning Company.

Auguste, B., Kihn, P., & Miller, M. (2010, September). *Closing the talent gap: Attracting and retaining top third graduates to a career in teaching* (Rep.). Retrieved from http://mckinseyonsociety.com/closing-the-talent-gap/

Aviv, R. (2014, July 21). Wrong answer: In an era of high-stakes testing, a struggling school made a shocking choice. *The New Yorker*. Retrieved from http://www.newyorker.com/magazine/2014/07/21/wrong-answer

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., . . . Shepard, L. A. (2010, August). *Problems with the use of student test scores to evaluate teachers* (Rep.). Retrieved from http://www.epi.org/files/page/-/pdf/bp278.pdf

Ballou, D. (2005). *Value-added assessment: Lessons from Tennessee* (Rep.). Retrieved from http://www.dpi.state.nc.us/docs/superintendents/quarterly/2010-11/20100928/ballou-lessons.pdf

Ballou, D., Mokher, C. G., & Cavalluzzo, L. (2012, March 1). *Using value-added assessment for personnel decisions: How omitted variables and model specification influence teachers' outcomes* (Rep.). Retrieved from https://aefpweb.org/sites/default/files/webform/AEFP-Using%20VAM%20for%20personnel%20decisions_02-29-12.pdf

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavior Statistics, 29*(1), 37-65. Retrieved from http://www.jstor.org/stable/3701306?seq=1#page_scan_tab_contents

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191-215. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/847061

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: Worth.

Batten, D. (2013). *Teacher performance evaluations and value-added scores: Evidence from North Carolina public schools* (Unpublished master's thesis). University of North Carolina at Chapel Hill.

Batten, D., Britt, C., DeNeal, J., & Hales, L. (2012, August). *NC teacher evaluations & teacher effectiveness: Exploring the relationship between value-added data and teacher evaluations* (Rep.). Retrieved from http://www.dpi.state.nc.us/docs/intern-research/reports/teachereval.pdf

Baxter, P., & Jack, S. (2008). Qualitative case study methodology: Study design and implementation for novice researchers. *The Qualitative Report, 13*(4), 544-559. Retrieved from http://www.nova.edu/ssss/QR/QR13-4/baxter.pdf

Berliner, D. (2009, March). *Poverty and potential: Out-of-school factors and school success* (Rep.). Retrieved from http://nepc.colorado.edu/files/PB-Berliner-NON-SCHOOL.pdf

Berman, E. (2003, March). *A short guide to evaluating teaching*. (Rep.). University of Arizona: Assessment and Enrollment Research. Retrieved from http://www.tlataskforce.uconn.edu/docs/resources/A_Short_Guide_to_Evaluating_Teaching.pdf

Berry, B., Daughtrey, A., & Wieder, A. (2009, December). *Collaboration: Closing the effective teaching gap* (Rep.). Retrieved from http://files.eric.ed.gov/fulltext/ED509717.pdf

Bianchi, A. (2003). A new look at accountability: "Value-Added" assessment. *Forecast - Emerging Issues in Public Education, 1*(1), 1-4. Retrieved from http://www.nyssba.org/clientuploads/forecast%20pdf/forecast0603.pdf

Bill and Melinda Gates Foundation. (2010). Measures of effective teaching. Retrieved from http://www.arteducators.org/news/Press_Release_Gates_Teacher_Evaluation_1-8-13.pdf

Bill and Melinda Gates Foundation. (2012). *Primary sources: 2012 America's teachers on the teaching profession* (Rep.). Retrieved from http://www.scholastic.com/primarysources/pdfs/Gates2012_full.pdf

Blase, J., & Blase, J. (2000). Effective instructional leadership: Teachers' perspectives on how principals promote teaching and learning in schools. *Journal of Educational Administration, 38*(2), 130-141. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.456.5752&rep=rep1&type=pdf

Bock, R. D., Wolfe, R., & Fisher, T. H. (1996). *A review and analysis of the Tennessee value-added assessment system* (Rep.). Retrieved from http://www.cgp.upenn.edu/pdf/Bock-Review_of_TVASS.PDF

Boser, U. (2012, March). *Race to the top: What have we learned from the states so far?* (Rep.). Retrieved from https://www.americanprogress.org/wp-content/uploads/issues/2012/03/pdf/rtt_states.pdf

Bowers, M. J., Wilson, R. E., & Hyde, R. L. (2011, June 30). *Criterion-Referenced Competency Test (CRCT) Report - Volume 3* (State of Georgia, Georgia Bureau of Investigation, Office of State Inspector General). Retrieved from http://www.atlanta.k12.ga.us/Page/410

Brandt, C., Mathers, C., Oliva, M., Brown-Sims, M., & Hess, J. (2007, November). *Examining district guidance to schools on teacher evaluation policies in the midwest region* (Rep.). Retrieved from http://ies.ed.gov/ncee/edlabs/regions/midwest/pdf/rel_2007030.pdf

Braskamp, L. A. (1980). What research says about the components of teaching. In W. R. Duckett (Ed.), *Observation and the evaluation of teaching (pp. 62-86)*. Bloomington, IN: Phi Delta Kappa.

Braun, H. (2005, September). *Using student progress to evaluate teachers: A primer on value-added models* (Rep.). Retrieved from http://www.ets.org/Media/Research/pdf/PICVAM.pdf

Braun, H., Chudowsky, N., & Koenig, J. (2010). *Getting value out of value-added report of a workshop* (Rep.). Retrieved from http://www.nap.edu/catalog/12820/getting-value-out-of-value-added-report-of-a-workshop

Brezicha, K., Bergmark, U., & Mitra, D. L. (2014). One size does not fit all: Differentiating leadership to support teachers in school reform. *Educational Administration Quarterly, 51*(1), 96-132. doi:10.1177/0013161X14521632

Briggs, D., & Domingue, B. (2011, February). *A review of the value-added analysis underlying the effectiveness rankings of Los Angeles Unified School District teachers by the Los Angeles Times* (Rep.). Retrieved from http://nepc.colorado.edu/files/NEPC-RB-LAT-VAM_0.pdf

Bryant, C. B. (2013). *Perceptions of Tennessee school principals about the Tennessee Educator Acceleration Model (TEAM)* (Doctoral dissertation, East Tennessee State University, 2013) (pp. 1-144). Johnson City, TN: East Tennessee State University Digital Commons.

Bryman, A. (2001). *Social research methods*. Oxford: Oxford University Press.

Campbell, D. (1976, December). *Assessing the impact of planned social change* (Rep.). Retrieved from http://portals.wi.wur.nl/files/docs/ppme/Assessing_impact_of_planned_social_change.pdf

Ceperley, P. E., & Reel, K. (1997). The impetus for the Tennessee value-added accountability system. In J. Millman (Author), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 133-136). Thousand Oakes, CA: Corwin Press.

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. London: Sage Publications.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011, December). The long-term impacts of teachers: Value-added and student outcomes in adulthood. Retrieved from http://news.harvard.edu/gazette/story/2012/02/superstar-teachers/

Cody, C. A., McFarland, J., Moore, J., & Preston, J. (2010, August). *The evolution and use of growth models* (Rep.). Retrieved from http://www.ncpublicschools.org/docs/intern-research/reports/growth.pdf

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6th ed.). London: RoutledgeFalmer.

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

Collins, C. (2014). Houston, we have a problem: Teachers find no value in the SAS education Value-Added Assessment System (EVAAS®). *Education Policy Analysis Archives, 22*(98), 1-42. Retrieved from file:///Users/test/Downloads/1594-6361-1-PB%20(4).pdf

Complete College America. (2012, April). *Remediation: Higher education's bridge to nowhere* (Rep.). Retrieved from http://completecollege.org/docs/CCA-Remediation-final.pdf

The Conference Board, Partnership for 21st Century Skills, Corporate Voices for Working Families, & Society for Human Resource Management. (2006, October). *Are they really ready to work?* (Rep.). Retrieved from https://www.conference-board.org/publications/publicationdetail.cfm?publicationid=1218

Cook, M. A., & Richards, H. C. (1972). Dimensions of principal and supervisor ratings of teacher behavior. *Journal of Experimental Education, 41*(2), 11-14.

Cooper, H. M. (1988). Organizing knowledge synthesis: A taxonomy of literature reviews. *Knowledge in Society, 1*, 104-126.

Creswell, J. W. (2003). *Research design: Qualitative, quantitative, and mixed method approaches*. Thousand Oaks, CA: Sage Publications.

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Upper Saddle River, NJ: Pearson/Merrill Prentice Hall.

Curtis, R. (2011, March). *District of Columbia Public Schools: Defining instructional expectations and aligning accountability and support* (Rep.). Retrieved from http://www.aspeninstitute.org/sites/default/files/content/docs/education/AI_DCPS_teacher%20evaluation.pdf

Curtis, R., & Wiener, R. (2012, January). *Means to an end: A guide to developing teacher evaluation systems that support growth and development* (Rep.). Retrieved from http://www.aspendrl.org/portal/browse/DocumentDetail?documentId=1008&download

Daley, G., & Kim, L. (2010, August). *A teacher evaluation system that works* (Rep.). Retrieved from http://tapsystem.niet.org/publications/rb_eval.pdf

Dancey, C. P., & Reidy, J. (2004). *Statistics without maths for psychology: Using SPSS for Windows*. New York: Prentice Hall.

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

Danielson, C. (2010/2011). Evaluations that help teachers learn. *Educational Leadership, 68*(4), 35-39.

Danielson, C., & McGreal, T. L. (2000). *Teacher evaluation to enhance professional practice*. Alexandria, VA: Association for Supervision and Curriculum Development.

Darling-Hammond, L. (2010, October). *Evaluating teacher effectiveness: How teacher performance assessments can measure and improve teaching* (Rep.). Retrieved from https://edpolicy.stanford.edu/sites/default/files/publications/evaluating-teacher-effectiveness_0.pdf

Darling-Hammond, L., & Adamson, F. (2013). *Developing assessments of deeper learning: The costs and benefits of using tests that help students learn* (Rep.). Retrieved from https://edpolicy.stanford.edu/sites/default/files/publications/developing-assessments-deeper-learning-costs-and-benefits-using-tests-help-students-learn_1.pdf

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E. H., & Rothstein, J. (2011, September 14). *Getting teacher evaluation right: A background paper for policy makers* (Rep.). Retrieved from http://tx.aft.org/files/gettingteacherevaluationright.pdf

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan, 93*(6), 8-15. Retrieved from http://www.pdkmembers.org/members_online/members/orders.asp?action=results&t=A&desc=evaluating+teacher+evaluation&text=&lname_1=&fname_1=&lname_2=&fname_2=&kw_1=&kw_2=&kw_3=&kw_4=&mn1=&yr1=&mn2=&yr2=&c1=

Darling-Hammond, L., Cook, C., Jaquith, A., & Hamilton, M. (2012, May). *Creating a comprehensive system for evaluating and supporting effective teaching* (Rep.). Retrieved from https://edpolicy.stanford.edu/sites/default/files/publications/creating-comprehensive-system-evaluating-and-supporting-effective-teaching_1.pdf

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

Day, J. C., & Newburger, E. C. (2002, July). *The big payoff: Educational attainment and synthetic estimates of work-life learning* (Rep.). Retrieved from https://www.census.gov/prod/2002pubs/p23-210.pdf

Denzin, N. K. (1989). *The research act: A theoretical introduction to sociological methods*. Englewood Cliffs, NJ: Prentice Hall.

Dillon, S. (2011, December 15). Failure rate of schools overstated, study says. Retrieved from http://www.nytimes.com/2011/12/15/education/education-secretary-overstated-failing-schools-under-no-child-left-behind-study-says.html

Donaldson, M. L., & Peske, H. G. (2010, March 10). *Supporting effective teaching through teacher evaluation: A study of teacher evaluation in five charter schools* (Rep.). Retrieved from https://cdn.americanprogress.org/wp-content/uploads/issues/2010/03/pdf/teacher_evaluation.pdf

DuFour, R., & Marzano, R. J. (2009). High-leverage strategies for principal leadership. *Educational Leadership, 66*(5), 62-68. Retrieved from http://www.ascd.org/publications/educational_leadership/feb09/vol66/num05/High-Leverage_Strategies_for_Principal_Leadership.aspx

Duncan, A. (2009, July 2). Speech presented at Partners in Reform: Remarks of Arne Duncan to the National Education Association. Retrieved from http://www2.ed.gov/news/speeches/2009/07/07022009.html

Duncan, A. (2010, August 29). Let's unleash all data on teachers: U.S. secretary of education asks, what's there to hide? Retrieved from http://www.nydailynews.com/opinion/unleash-data-teachers-u-s-secretary-education-asks-hide-article-1.200867

Duncan, A. (2011a, August 17). College & career readiness improving among U.S. high school graduates. Retrieved from http://www.act.org/newsroom/college-career-readiness-improving-among-u-s-high-school-graduates/

Duncan, A. (2011b, March 9). Duncan says 82 percent of America's schools could "fail" under NCLB this year. Retrieved from http://www.ed.gov/news/press-releases/duncan-says-82-percent-americas-schools-could-fail-under-nclb-...

Duncan, A., & Obama, B. (2009, July 24). *President Obama and Secretary Duncan announce Race to the Top*. Speech presented at Press Release, Washington D.C. Retrieved from http://www.ed.gov/blog/2009/07/president-obama-secretary-duncan-announce-race-to-the-top/

Eaker, R. E., DuFour, R., & DuFour, R. B. (2002). *Getting started: Reculturing schools to become professional learning communities*. Bloomington, IN: National Educational Service.

Eckert, J. M., & Dabrowski, J. (2010, May). Should value-added measures be used for performance pay? *Kappan, 91*(8), 88-92. Retrieved from http://projects.brevardschools.org/PAS/Shared%20Documents/Shouldvalueaddedmeasures.pdf

Eichenwald, K. (2012, August). Microsoft's lost decade. *Vanity Fair*. Retrieved from http://www.vanityfair.com/news/business/2012/08/microsoft-lost-mojo-steve-ballmer

Eisner, E. (1982). An artistic approach to supervision. In T. Sergiovanni (Ed.), *Supervision of teaching: 1982 ASD yearbook*. Alexandria, VA: Association for Supervision and Curriculum Development.

Elmore, R. F. (2000). *Building a new structure for school leadership*. Washington, DC: Albert Shanker Institute.

Federal Education Policy History. (2011, February). Elementary and Secondary Act of 1965. Retrieved from https://federaleducationpolicy.wordpress.com/2011/02/19/1965-elementary-and-secondary-education-act/

Feng, L., Figlio, D. N., & Sass, T. (2010, June). *School accountability and teacher mobility* (Rep.). Retrieved from http://www.nber.org/papers/w16070.pdf

Fernandez, W. D. (2005). *The grounded theory method and case study data in IS research: Issues and design found in Information systems foundation: Constructing and criticizing* (D. N. Hart & S. D. Gregor, Eds.). Canberra, Australia: The Australian National University.

Fink, E., & Resnick, L. B. (2001). Developing principals as instructional leaders. *Phi Delta Kappan, 82*(8), 598-606. Retrieved from http://www.easybib.com/journal-article-citation/from-pubdatabase

Firestone, W. A., Monfils, L., & Schorr, R. Y. (2004). Test preparation in New Jersey: Inquiry-oriented and didactic responses. *Assessment in Education: Principles, Policy & Practice, 11*(1), 67-88. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/0969594042000209001?journalCode=caie20

Fleischman, H. L., Hopstock, P. J., Pelczar, M. P., Shelley, B. E., & Xie, H. (2010, December). Highlights from PISA 2009: Performance of U.S. 15-year-old students in reading, mathematics, and science literacy in an international context. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2011004

Flick, A. (2009). *An introduction to qualitative research*. Thousand Oaks, CA: Sage Publications.

Frase, L. E. (1992). Constructive feedback on teaching is missing. *Education, 113*(2), 176-181. Retrieved from http://connection.ebscohost.com/c/articles/9303020706/constructive-feedback-teaching-missing

Frase, L., Downey, C., & Canciamilla, L. (1999). Putting principals in their place: The classroom. *Thrust for Educational Leadership, 28*(5), 36-39. Retrieved from http://connection.ebscohost.com/c/articles/2134860/putting-principals-their-place-classroom

Frase, L. E., & Streshly, W. (1994). Lack of accuracy, feedback, and commitment in teacher evaluation. *Journal of Personnel Evaluation in Education, 8*(1), 47-57. doi:10.1007/BF00972709

Fuhrman, S. H., & Elmore, R. F. (2004). *Redesigning accountability systems for education*. New York, NY: Teachers College Press.

Gable, R. K., & Wolf, M. B. (1993). *Instrument development in the affective domain: Measuring attitudes and values in corporate and school settings* (2nd ed.). Norwell, MA: Kluwer Academic.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Education research: An introduction* (6th ed.). White Plains, NY: Longman.

Gardner, D. P. (1983, April). *A nation at risk: The imperative for educational reform* (Rep.). Retrieved from http://datacenter.spps.org/uploads/sotw_a_nation_at_risk_1983.pdf

Gay, L. R., Mills, G. E., & Airasian, P. W. (2009). *Educational research: Competencies for analysis and applications*. Upper Saddle River, NJ: Merrill/Prentice Hall.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference* (4th ed.). Boston: Allyn and Bacon.

Gerring, J. (2007). *Case study research: Principles and practices*. New York, NY: Cambridge University Press.

Gill, B., Bruch, J., & Booker, K. (2013, September 13). *Using alternative student growth measures for evaluating teacher performance: What the literature says* (Rep.). Retrieved http://files.eric.ed.gov/fulltext/ED544205.pdf

Gillum, J., & Bello, M. (2011, March 30). When standardized test scores soared in D.C., were the gains real? - USATODAY.com. Retrieved from http://usatoday30.usatoday.com/news/education/2011-03-28-1Aschooltesting28_CV_N.htm

Glaser, B. (1998). *Doing Grounded theory: Issues and discussion*. Mill Valley, CA: Sociology Press.

Glazerman, S., Chiang, H., Wellington, A., Constantine, J., & Player, D. (2011, October). *Impacts of performance pay under the teacher incentive fund: Study design report* (Rep.). Retrieved from http://www.mathematica-mpr.com/~/media/publications/PDFs/education/performpay_TIF.pdf

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G. (2010, November 17). *Evaluating teachers: The important role of value-added* (Rep.). Retrieved from http://www.brookings.edu/~/media/research/files/reports/2010/11/17-evaluating-teachers/1117_evaluating_teachers.pdf

Glazerman, S., Protik, A., Teh, B., Bruch, J., & Max, J. (2013, November). *Transfer incentives for high-performing teachers: Final results from a multisite randomized experiment* (Rep.). Retrieved from https://ies.ed.gov/ncee/pubs/20144003/pdf/20144003.pdf

Goe, L. (2008, May). *Key issue: Value-added models to identify and support highly effective teachers* (Rep.). Retrieved from http://www.gtlcenter.org/sites/default/files/docs/strategies/het/UsingValueAddedModels.pdf

Goldhaber, D. (2002). The mystery of good teaching. *Education Next*. Retrieved from http://educationnext.org/files/ednext20021_50.pdf

Goldhaber, D., & Hansen, M. (2010, February). *Assessing the potential of using value-added estimates of teacher performance for making tenure decisions* (Working paper No. 31). Retrieved from http://www.urban.org/sites/default/files/alfresco/publication-pdfs/1001369-Assessing-the-Potential-of-Using-Value-Added-Estimates-of-Teacher-Job-Performance-for-Making-Tenure-Decisions.PDF

Goldhaber, D., & Theobald, R. (2013). *Do different value-added models tell us the same thing?* (Rep.). Retrieved from http://www.carnegieknowledgenetwork.org/wp-content/uploads/2012/10/CKN_2012-10_Goldhaber_Nov2013-Update.pdf

Goldstein, J. (2004). Making sense of distributed leadership: The case of peer assistance and review. *Educational Evaluation and Policy Analysis, 26*(2), 173-197. Retrieved from http://www.isbe.net/racetothetop/PDF/par-goldstein-2004.pdf

Goldstein, J. (2007). Easy to dance to: Solving the problems of teacher evaluation with peer assistance and review. *American Journal of Education, 113*(3), 479-508. Retrieved from http://www.isbe.state.il.us/racetothetop/pdf/par-goldstein-2007.pdf

Goldstein, J., & Noguera, P. A. (2006). A thoughtful approach to teacher evaluation. *Educational Leadership, 63*(6), 31-37. Retrieved from http://www.ascd.org/publications/educational-leadership/mar06/vol63/num06/A-Thoughtful-Approach-to-Teacher-Evaluation.aspx

Gonzales, P., Williams, T., Jocelyn, L., Roey, S., Kastberg, D., & Brenwald, S. (2009, September). Highlights from TIMSS 2007: Mathematics and science achievement of U.S. fourth- and eighth-grade students in an international context. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009001

Gonzalez, T. (2012, July 17). TN education reform hits bump in teacher evaluation. Retrieved from http://www.wbir.com/news/article/226990/0/TN-education-reform-hits-bump-in-teacher-evaluation

Good, T. L., & Brophy, J. E. (1984). *Looking in classrooms*. New York: Harper & Row.

Good, T. L., & Duckett, W. R. (1980). *Classroom observations: Potential and problems*. Bloomington, IN: Phi Delta Kappa, Center on Evaluation, Development and Research.

Goodlad, J. I., & Klein, M. F. (1970). *Looking behind the classroom door: A useful guide to observing schools in action*. Worthington, OH: C. A. Jones Pub.

Gordon, R., Kane, T. J., & Staiger, D. O. (2006, April). *Identifying effective teachers using performance on the job* (Rep.). Retrieved from http://www.brookings.edu/research/papers/2006/04/education-gordon

Government Accountability Office. (2009, July). Teacher quality: Sustained coordination among key federal education programs could enhance state efforts to improve teacher quality. Retrieved from http://www.gao.gov/products/GAO-09-593

Government Accountability Office. (2013, May). *K-12 education: States' test security policies and procedures varied* (Rep.). Retrieved from http://www.gao.gov/assets/660/654721.pdf

Green, J., & Wintfield, N. (1995). Report cards on cardiac surgeons: Assessing New York State's approach. *New England Journal of Medicine, 332*, 1229-1232. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7700321

Griffith, F. J. (1973). *A handbook for the observation of teaching and learning*. Midland, MI: Pendell Pub.

Guarino, C., & Stacy, B. (2012, March 13). *Review of gathering feedback for teaching* (Rep.). Retrieved from http://nepc.colorado.edu/files/TTR-MET-Gates_0.pdf

Haertel, E. H. (2013). *Reliability and validity of inferences about teacher based on student test scores* (Rep.). Retrieved from https://www.ets.org/Media/Research/pdf/PICANG14.pdf

Hamre, B. K., Pianta, R. C., Mashburn, A. J., & Downer, J. T. (2007, April 1). *Building a science of classrooms: Application of the CLASS framework in over 4,000 U.S. early childhood and elementary classrooms* (Rep.). Retrieved from http://fcd-us.org/sites/default/files/BuildingAScienceOfClassroomsPiantaHamre.pdf

Hanover Research. (2011, June). A survey of Race to the Top teacher evaluation systems. Retrieved from http://www.hanoverresearch.com/k-12-sample-reports/

Hanover Research. (2012, March). *Best practices for including multiple measures in teacher evaluations* (Rep.). Retrieved from http://www.hanoverresearch.com/wp-content/uploads/2012/05/Best-Practices-for-Including-Multiple-Measures-in-Teacher-Evaluations-Membership.pdf

Hanushek, E. A. (1972). *Education and race; an analysis of the educational production process*. Lexington, MA: Lexington Books.

Harris, B. M. (1986). *Developmental teacher evaluation*. Boston: Allyn and Bacon.

Harris, P., Smith, B. M., & Harris, J. (2011). *The myths of standardized testing: Why they don't tell you what you think they do*. New York, NY: Rowman & Littlefield.

Headden, S. (2011). *Inside IMPACT: D.C.'s model teacher evaluation system* (Rep.). Retrieved from http://www.nnstoy.org/download/evalution/Impact%20Report%20Release.pdf

Heneman III, H. G., Milanowski, A., Kimball, S. M., & Odden, A. (2006, May). *Standards-based teacher evaluation as a foundation for knowledge-and skill-based pay* (Rep.). Retrieved from http://cpre.wceruw.org/publications/rb45.pdf

Henry, A. E. (2011). *Advantages to and challenges of using ratings of observed teacher-child interactions* (Doctoral dissertation, University of Virginia, 2010) (pp. 1-179). Ann Arbor, MI: Proquest.

Hershberg, T. (2005). Value-added assessment and systemic reform: A response to the challenge of human capital development. *Phi Delta Kappa, 87*, 276-283.

Ho, A. D., & Kane, T. J. (2013, January). *The reliability of classroom observations by school personnel* (Rep.). Retrieved from http://www.metproject.org/downloads/MET_Reliability_of_Classroom_Observations_Research_Paper.pdf

Hoffman, D. A., Jacobs, R., & Gerras, S. J. (1992). Mapping individual performance over time. *Journal of Applied Psychology, 77*, 185-195. Retrieved from http://psycnet.apa.org/index.cfm?fa=buy.optionToBuy&id=1992-29427-001

Holtzapple, E. (2003). Criterion-related validity evidence for a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education, 17*(3), 207-219.

Huffman, J. B., & Hipp, K. K. (2003). *Reculturing schools as professional learning communities*. Lanham, MD: Scarecrow Education.

Hull, J. (2011, March 31). Building a better evaluation system: At a glance. Retrieved from http://www.centerforpubliceducation.org/Main-Menu/Staffingstudents/Building-A-Better-Evaluation-System/default.aspx

Hull, J. (2013, October). *Trends in teacher evaluation: How states are measuring teacher performance* (Rep.). Retrieved from http://www.centerforpubliceducation.org/Main-Menu/Evaluating-performance/Trends-in-Teacher-Evaluation-At-A-Glance/Trends-in-Teacher-Evaluation-Full-Report-PDF.pdf

Humphrey, D. C., Koppich, J. E., Bland, J. A., & Bosetti, K. R. (2011). *Peer review: Getting serious about teacher support and evaluation* (Rep.). Retrieved from https://www.sri.com/sites/default/files/publications/par-report-2011-final-sept6.pdf

Institute for Educational Leadership (IEL). (2001, April). *Leadership for student learning: Redefining the teacher as leader* (Rep.). Retrieved from http://iel.org/sites/default/files/Leadership-for-Student-Learning-Series-2-Teacher-04-2001.pdf

Jacob, B. A., & Lefgren, L. (2008). *Can principals identify effective teachers? Evidence on subjective performance evaluation in education* (Rep.). Retrieved from http://econpapers.repec.org/article/ucpjlabec/v_3a26_3ay_3a2008_3ap_3a101-136.htm

Jacques, C. (2013, May). *Leveraging teacher talent: Peer observation in educator evaluation* (Rep.). Retrieved from http://www.gtlcenter.org/sites/default/files/docs/GTL_AskTeam_LeveragingTeacherTalents.pdf

Jerald, C. (2012, March). *Ensuing accurate feedback from observations - Perspectives on practice* (Rep.). Retrieved from https://docs.gatesfoundation.org/Documents/ensuring-accuracy-wp.pdf

Joe, J., Tocci, C. M., Holtzman, S. L., & Williams, J. C. (2013, June). *Foundations of observation considerations for developing a classroom observation system that helps districts achieve consistent and accurate scores* (Rep.). Retrieved from http://www.metproject.org/downloads/MET-ETS_Foundations_of_Observation.pdf

Johnson, M., Lipscomb, S., & Gill, B. (2013, November). *Sensitivity of value-added teacher estimates to student and peer control variables* (Rep.). Retrieved from http://www.mathematica-mpr.com/~/media/publications/PDFs/education/value-added_estimates_wp.pdf

Jones, B. D., & Egley, R. J. (2007). Learning to take tests or learning for understanding? Teachers' beliefs about test-based accountability. *The Educational Forum, 71*, 232-248. Retrieved from http://files.eric.ed.gov/fulltext/EJ763214.pdf

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013, January). *Have we identified effective teachers: Validating measures of effective teaching using random assignment* (Rep.). Retrieved from http://www.metproject.org/downloads/MET_Validating_Using_Random_Assignment_Research_Paper.pdf

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (Working paper). Retrieved from http://www.nber.org/papers/w14607.pdf

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2010, March). *Identifying effective classroom practices using student achievement data* (Working paper). Retrieved from http://www.serve.org/uploads/docs/Events%20&%20Webinars/CPS_TES_073010_JHR_Revise_final.pdf

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011, Summer). Evaluating teacher effectiveness. *Education Next,* 54-60. Retrieved from http://educationnext.org/files/ednext_20113_research_kane.pdf

Kaye, R. (2010, February 24). All teachers fired at Rhode Island school. Retrieved from http://www.cnn.com/2010/US/02/24/rhode.island.teachers/

Koedel, C., & Betts, J. R. (2007, April). *Re-examining the role of teacher quality in the educational production function* (Rep.). Retrieved from https://economics.missouri.edu/working-papers/2007/wp0708_koedel.pdf

Koedel, C., & Betts, J. (2009, March). *Value-added to what? How a ceiling in the testing instrument influences value-added estimation* (Rep.). Retrieved from http://www.nber.org/papers/w14778.pdf

Koppich, J. E. (2009, October). *Strategic management of human capital - Toledo: Peer assistance and review (PAR)* (Rep.). Retrieved from file:///Users/test/Downloads/SMHC_Toledo_case%20(1).pdf

Korb, K. A. (n.d.). *Calculating reliability of quantitative measures* (Rep.). Retrieved from http://korbedpsych.com/LinkedFiles/CalculatingReliability.pdf

Koretz, D. M. (2008). *Measuring up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee value added assessment system. *Educational Evaluation and Policy Analysis, 25*(3), 287-298. Retrieved from http://www.cgp.upenn.edu/pdf/Kuppermintz_-_Teacher_Effects.pdf

Lachat, M. A. (2001). *Data-driven high school reform: The breaking ranks model* (Rep.). Retrieved from http://www.brown.edu/academics/education-alliance/sites/brown.edu.academics.education-alliance/files/publications/datdrv_hsrfm.pdf

Ladd, H. F., & Walsh, R. P. (2002). Implementing value-added measures of school effectiveness: Getting the incentives right. *Economics of Education Review, 21*(1), 1-17. Retrieved from http://www.sciencedirect.com/science/article/pii/S027277570000039X

Ladner, M., Lips, D., & Daniels, M. (2012). *ALEC report card on American education* (Rep.). Retrieved from http://www.alec.org/docs/17thReportCard/ALECs_17th_Report_Card.pdf

Lauen, D. L., Henry, G. T., Rose, R. A., & Kozlowski, K. P. (2013, December). *The distribution of teacher value added in North Carolina* (Rep.). Retrieved from http://cerenc.org/wp-content/uploads/2013/12/Baseline-TQ-Report_FINAL_12-05-2013.pdf

Lehmann, H. (2001). *A grounded theory of international information systems* (Doctoral dissertation, The University of Auckland) (pp. 1-291). The University of Auckland Library. Retrieved from https://researchspace.auckland.ac.nz/handle/2292/626

Levin, B. (1979, December). *Teacher evaluation - A review of research* (Rep.). Retrieved from http://www.ascd.org/ASCD/pdf/journals/ed_lead/el_197912_levin.pdf

Little, O., Goe, L., & Bell, C. (2009, April). *A practical guide to evaluating teacher effectiveness* (Rep.). Retrieved from http://www.gtlcenter.org/sites/default/files/docs/practicalGuide.pdf

Lockwood, J., Louis, T. A., & McCaffrey, D. F. (2002). Uncertainty in rank estimation: Implications for value-added modeling accountability systems. *Journal of Educational and Behavioral Statistics, 27*(3), 255-270. doi:10.3102/10769986027003255

Lohman, J. (2010, June 4). Comparing the No Child Left Behind Act and the Race to the Top. Retrieved from http://www.cga.ct.gov/2010/rpt/2010-R-0235.htm

Louis, K. S., Leithwood, K., Wahlstrom, K. L., & Anderson, S. E. (2010). *Investigating the links to improved student learning: Final report of research findings* (Rep.). Retrieved from http://www.wallacefoundation.org/knowledge-center/school-leadership/key-research/Documents/Investigating-the-Links-to-Improved-Student-Learning.pdf

Lynn, H., Barrett, N., Marks, J., Comperatore, A., Henry, G. T., & Guthrie, J. (2013, September). *Measures of student growth in the North Carolina Educator Evaluation System* (Rep.). Retrieved from http://cerenc.org/wp-content/uploads/2011/10/FINAL-Formative-report-TLEE-measures-of-student-growth-9-x-13.pdf

MacNeill, N. (2007). Pedagogic obsolescence: A curtain call for school principalship. In *Australian Association for Research in Education (AARE) Conference*. Retrieved from http://www.aare.edu.au/data/publications/2007/mac07041.pdf

Markley, T. (2006). *Defining the effective teacher: Current arguments in education* (Rep.). Retrieved from http://www.ibrarian.net/navon/page.jsp?paperid=882968&searchTerm=act+lld+lld

Marshall, K. (2009). *Rethinking teacher supervision and evaluation: How to work smart, build collaboration, and close the achievement gap*. San Francisco, CA: Jossey-Bass.

Marzano Research. (2011, March 1). What works in Oklahoma schools: A comprehensive needs assessment of Oklahoma schools. Retrieved from http://www.marzanoresearch.com/teacher-effectiveness/what-works-in-oklahoma-schools-a-comprehensive-needs-assessment-of-oklahoma-schools

Marzano, R., Toth, M., & Schooling, P. (n.d.). *Contemporary research base for the Marzano causal teacher evaluation model* (Rep.). Retrieved from http://www.marzanocenter.com/files/MC-White-Paper-20120605.pdf

Marzano, R. J., Waters, T., & McNulty, B. A. (2005). *School leadership that works: From research to results*. Alexandria, VA: Association for Supervision and Curriculum Development.

McCaffrey, D. F. (2013, June). *Do value-added methods level the playing field for teachers?* (Rep.). Retrieved from http://www.carnegieknowledgenetwork.org/wp-content/uploads/2013/06/CKN_2012-10_McCaffrey.pdf

McCaffrey, D. F., & Lockwood, J. (2008, November 13). *Value-added models: Analytic issues* [PDF]. Arlington, VA: The Rand Corporation.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67-101. doi:10.3102/10769986029001067

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *American Education Finance Association, 4*(4), 572-606. Retrieved from http://www.utla.net/system/files/mccaffrey_study.pdf

McCullough, L. (2011, October 20). *Supporting rater accuracy and consistency in classroom observations* (Powerpoint). Retrieved from http://cepr.harvard.edu/files/cepr/files/ncte-webinar-teachstone-powerpoint.pdf

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

McIntyre, D. J. (1980). Teacher evaluation and the observer effect. *NASSP Bulletin, 64*(434), 36-40. doi:10.1177/019263658006443408

McLaughlin, M. W., & Pfeifer, R. S. (1988). *Teacher evaluation: Improvement, accountability, and effective learning*. New York, NY: Teachers College Press.

McNeil, P. W. (2003, January). *Rethinking high school: The next frontier for state policymakers* (Rep.). Retrieved from http://www.aspeninstitute.org/publications/rethinking-high-school-next-frontier-state-policymakers

Measures of Effective Teaching. (2010, December). *Learning about teaching: Initial findings from the Measures of Effective Teaching project* (Rep.). Retrieved from http://metproject.org/reports.php

Measures of Effective Teaching. (2012, January). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains* (Rep.). Retrieved from http://www.metproject.org/reports.php

Measures of Effective Teaching. (2013, January). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study* (Rep.). Retrieved from http://metproject.org/reports.php

Medley, D. M., & Coker, H. (1987). The accuracy of principals' judgments of teacher performance. *Journal of Educational Research, 80*(4), 242-247.

Metlife. (2012, March). *The MetLife survey of the American teacher* (Rep.). Retrieved from http://files.eric.ed.gov/fulltext/ED530021.pdf

Miami-Dade County Public Schools. (2015). *Instructional performance evaluation and growth system (IPEGS) - Procedural handbook 2015 edition* (Rep.). Retrieved from http://ipegs.dadeschools.net/pdfs/2011-2012_IPEGS_Procedural_Handbook.pdf

Mid-continent Research for Education and Learning, North Carolina Department of Instruction, & North Carolina State Board of Education. (2012, September). *North Carolina teacher evaluation process* [PDF]. Denver: Mid-continent Research for Education and Learning.

Milanowski, A. T., Heneman III, H. G., & Kimball, S. M. (2011, March). *Teaching assessment for teacher human capital management: Learning from the current state of the art* (Working paper). Retrieved from http://www.wcer.wisc.edu/publications/workingpapers/Working_Paper_No_2011_02.pdf

Milanowski, A. T., Kimball, S. M., & White, B. (2004, March). *The relationship between standards-based teacher evaluation scores and student achievement: Replication and extensions at three sites* (Working paper). Retrieved from http://www.cpre.wceruw.org/papers/3site_long_TE_SA_AERA04TE.pdf

Milner, J. O. (February 1991). Suppositional style and teacher evaluation. *Phi Delta Kappa, 72(6)*, 464-467.

Murnane, R. J. (1975). *The impact of school resources on the learning of inner city children*. Cambridge, MA: Ballinger Pub.

Murphy, J. (2011, June 8). Virtually all of Pennsylvania teachers, principals rated satisfactory, but state education secretary says that doesn't square with student test scores | PennLive.com. Retrieved from http://www.pennlive.com/midstate/index.ssf/2011/06/all_but_1_percent_of_pennsylva.html

Murphy, D. (2012, April). *Where is the value in value-added modeling?* (Rep.). Retrieved from http://images.pearsonassessments.com/images/tmrs/Where_is_the_Value_in_Value-Added_Modeling.pdf

Nappi, J. S. (2014). The teacher leader: Improving schools by building social capital through shared leadership. *National Association of Secondary School Principals, 9*(6), 1-6. Retrieved from http://www.nassp.org/Content/158/November_2014_PRR.pdf

National Academy Press (NAP). (2009, October 5). *Letter report to the U.S. Department of Education on the Race to the Top fund* (Rep.). Retrieved from http://www.nap.edu/read/12780/chapter/1

National Association of Secondary Principals. (1996). *Breaking ranks: Changing an American institution* (Rep.). Retrieved from http://www.principals.org/Content.aspx?topic=47132

National Center for Education Statistics. (1999). *Teacher quality: A report on the preparation and quality of public school teachers*. Washington, DC: U.S. Department of Education Office of Educational Research and Improvement.

National Center for Education Statistics. (2011a). The nation's report card: Mathematics 2011. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2012458

National Center for Education Statistics. (2011b). The nation's report card: Reading 2011. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2012457

National Center for Education Statistics. (2011c). The nation's report card: Science 2011. Retrieved from http://nces.ed.gov/nationsreportcard/pubs/main2011/2012465.asp

National Center for Higher Education Management Systems. (2012). College participation rates: College-going rates of high school graduates - directly from high school. Retrieved from http://higheredinfo.org/dbrowser/index.php?submeasure=63

National Council on Teacher Quality. (2009, January). *A Race to the Top scorecard* (Rep.). Retrieved from http://www.nctq.org/dmsView/Race_to_the_Top_Scorecard_NCTQ_Report

National Council on Teacher Quality. (2011a, October). *State of the states: Trends and early lessons on teacher evaluation and effectiveness policies* (Rep.). Retrieved from http://www.edweek.org/media/nctq_stateofthestates-14hawaii.pdf

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

National Research Council. (1999). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Natriello, G. (1983, November). *Valuation frequency, teacher influence, and the internalization of evaluation processes: A review of six studies using the theory of evaluation and authority* (Rep.). Retrieved from http://files.eric.ed.gov/fulltext/ED242050.pdf (ERIC Document Reproduction Service)

New America. (2013, May). No child left behind. Retrieved from http://atlas.newamerica.org/no-child-left-behind-overview

The New Teacher Project. (2010). *Teacher evaluation 2.0* (Rep.). Retrieved from http://tntp.org/assets/documents/Teacher-Evaluation-Oct10F.pdf

Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Education Policy Analysis Archives, 18*(23). Retrieved from http://epaa.asu.edu/ojs/article/view/810/858

Nichols, S. L., Glass, G. V., & Berliner, D. C. (2005, September). *High stakes testing and student achievement: Problems for the No Child Left Behind Act* (Rep.). Retrieved from http://epsl.asu.edu/epru/documents/EPSL-0509-105-EPRU.pdf

North Carolina Department of Instruction. (2012, March). *Teacher effectiveness and support for growth: Using meaningful evaluation to increase effectiveness of teachers and leaders* (Rep.). Retrieved from http://www.ncpublicschools.org/docs/ready/resources/spring2012/teacher-effect.pdf

North Carolina Department of Instruction. (2013, December 10). *Measuring student learning for educator effectiveness: A guide to the use of student growth data in the evaluation of North Carolina teachers* (Rep.). Retrieved from http://www.dpi.state.nc.us/docs/effectiveness-model/student-growth/measuring-growth-guide.pdf

North Carolina Department of Instruction. (2014, June). Educator effectiveness data. Retrieved from http://www.dpi.state.nc.us/effectiveness-model/data/

Nussbaum, B. (2012, May 16). *The value-added model and its appropriate place in evaluating teachers* (Rep.). Retrieved from http://www.luc.edu/media/lucedu/law/centers/childlaw/childed/pdfs/2012studentpapers/nussbaum.pdf

Oakes, M., & Robertson, J. S. (2014, May). *North Carolina public school teachers: Reactions to teacher evaluations and merit pay* (Rep.). Retrieved from http://people.uncw.edu/robertsonj/documents/NC%20Teacher%20Evaluation%20and%20Merit%20Pay%20survey%20results%20Report.pdf?utm_content=buffer76bd2&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer

O'Donnell, P. (2013, June 18). Teachers' "value-added" ratings and relationship to student income levels questioned. Retrieved from http://www.cleveland.com/metro/index.ssf/2013/06/teachers_value-added_ratings_a.html#incart_story_package

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*(1), 163-193. Retrieved from http://aer.sagepub.com/content/48/1/163.abstract

Paulson, A. (2011, November 01). Education report card: Flat reading scores are "deeply disappointing." Retrieved from http://www.csmonitor.com/USA/Education/2011/1101/Education-report-card-Flat-reading-scores-are-deeply-disappointing

Pearson Learning. (2004, November). *Value-added assessment systems* (Rep.). Retrieved from http://images.pearsonassessments.com/images/tmrs/tmrs_rg/ValueAdded.pdf?WT.mc_id=TMRS_Value_Added

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

Piburn, M., & Swada, D. (2000). *Reformed teaching observation protocol (RTOP): Reference manual* (Rep.). Retrieved from https://mathed.asu.edu/instruments/rtop/RTOP_Reference_Manual.pdf

Popham, W. J. (1988). The dysfunctional marriage of formative and summative evaluation. *Journal of Personnel Evaluation in Education, 1*(3), 269-273. Retrieved from http://link.springer.com/article/10.1007%2FBF00123822?LI=true#page-1

Popham, W. J. (2013, March). *On serving two masters: Formative and summative evaluation* (Rep.). Retrieved from http://www.nassp.org/Content/158/PLmar13_popham.pdf

Preston, J. (2013, March 26). *North Carolina educator effectiveness model update* (Rep.). Retrieved from http://www.ncpublicschools.org/docs/ccsa/conference/2013/presentations/71.pdf

Prince, J. D. (1984). Formative teacher evaluation: The crucial element in an outcome-based education program. *Phi Delta Kappa: Evaluation of Teaching: The Formative Process,* 85-94.

Randolph, J. J. (2009). A guide to writing the dissertation literature review. *Practical Assessment, Research & Evaluation, 14*(13), 1-13.

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

Ravitch, D., & Rhee, M. (2011, August 18). *Separate but Unequal: Closing the education gap*. Speech presented at Closing the Education Gap - Moderated by Charlayne Hunter-Gault in Martha's Vineyard. Retrieved from http://dubois.fas.harvard.edu/video/separate-unequal-closing-education-gap-moderated-charlayne-hunter-gault

Raymond, A. A. (2007). Good teachers, or great? *Teaching Pre K-8, 37*(4), 6-7.

Reform Support Network. (n.d.). *Race to the top at a glance evaluations of teacher effectiveness: State requirements for classroom observations* (Rep.). Retrieved from https://www2.ed.gov/about/inits/ed/implementation-support-unit/tech-assist/evaluations-of-teacher-effectiveness.pdf

Ribas, W. B. (2002). *Teacher evaluation that works!! The educational, legal, public relations (political) and social-emotional (E.L.P.S.) standards and processes of effective supervision and evaluation*. Westwood, MA: Ribas Pub.

Ridgway, V. F. (1956). Dysfunctional consequences of performance measurements. *Administrative Science Quarterly, 1*(2), 240-247. Retrieved from https://www2.le.ac.uk/departments/management/documents/undergraduate-courses/a-f-preparatory-reading-ridgway.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and student achievement. *Econometrica, 73*(2), 417-458. Retrieved from http://www.econ.ucsb.edu/~jon/Econ230C/HanushekRivkin.pdf

Roberts, J. (2011, September 29). Teacher evaluations questioned as time-consuming. *Memphis Commercial Appeal*. Retrieved from http://wwwsecondchance.blogspot.com/2011/09/republicans-get-one-right.html

Rothstein, R. (2004, October). Class and the classroom. Retrieved from http://www.asbj.com/MainMenuCategory/Archive/2004/October/Class-and-the-Classroom.aspx

Rothstein, R. (2008, May). *Teacher quality in educational production: Tracking, decay, and student achievement* (Rep.). Retrieved from http://harris.princeton.edu/pubs/pdfs/25ers.pdf

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2003). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavior Statistics, 29*(1), 103-116. Retrieved from http://jeb.sagepub.com/content/29/1/103.extract

Sanders, W. L., & Horn, S. P. (1995). *The Tennessee value-added assessment system (TVAAS): Mixed model methodology in educational assessment* (Teacher evaluation: Guide to effective practice). Norwell, MA: Kluwer Academic.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee value-added assessment system (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*(3), 247-256.

Sanders, W. L., & Rivers, J. C. (1996, April). *Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research* (Rep.). Retrieved from https://www.sas.com/govedu/edu/ed_eval.pdf

Sanders, W. L., Wright, S. P., Rivers, J. C., & Leandro, J. G. (2009, November). *A response to criticisms of SAS® EVAAS®* (Rep.). Retrieved from https://www.sas.com/resources/asset/Response_to_Criticisms_of_SAS_EVAAS_11-13-09.pdf

Sarason, S. (1996). *Revisiting the culture of school and the problem of change*. New York, NY: Teachers College Press.

Sartain, L., Stoelinga, S. R., & Krone, E. (2011, November). Rethinking teacher evaluation: Findings from the first year of the Excellence in Teaching Project in Chicago public schools. Retrieved from https://ccsr.uchicago.edu/sites/default/files/publications/Teacher%20Eval%20Report%20FINAL.pdf

Sass, T. (2008, November). *The stability of value□added measures of teacher quality and implications for teacher compensation policy* (Rep.). Retrieved from http://www.caldercenter.org/sites/default/files/1001266_stabilityofvalue.pdf

Sass, T., Hannaway, J., Xu, Z., Figlio, D., & Feng, L. (2010, November). *Value-added of teachers in high-poverty schools and lower-poverty schools* (Rep.). Retrieved from http://www.urban.org/sites/default/files/alfresco/publication-pdfs/1001469-Value-Added-of-Teachers-in-High-Poverty-Schools-and-Lower-Poverty-Schools.PDF

Scherrer, J. (2012). What's the value of VAM (value-added modeling)? *Phi Delta Kappan, 93*(8), 58-60. Retrieved from http://pdk.sagepub.com/content/93/8/58.abstract

Schmoker, M. (2001). *The results fieldbook: Practical strategies from dramatically improved schools*. Alexandria, VA: Association for Supervision and Curriculum Development.

Schmoker, M. J. (2006). *Results now: How we can achieve unprecedented improvements in teaching and learning*. Alexandria, VA: Association for Supervision and Curriculum Development.

Schneider, M. (2009, Fall). *The international PISA test: A risky investment for states* (Rep.). Retrieved from http://educationnext.org/files/fall09-international-pisa.pdf

Schochet, P. Z., & Chiang, H. S. (2010, July). *Error rates in measuring teacher and school performance based on student test score gains* (Rep.). Retrieved from https://ies.ed.gov/ncee/pubs/20104004/pdf/20104004.pdf

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler & R. W. Gagne (Authors), *In American Educational Research Association Monograph Series on Curriculum Evaluation: Vol. 1. Perspectives of curriculum evaluation*. Chicago, IL: Rand McNally.

Senge, P. M., Kleiner, A., Roberts, C., & Roth, G. (1999). *The dance of change*. New York, NY: Doubleday.

Shepard, L. A. (1990, January). *"Inflated test scores gains": Is it old norms or teaching to the test?* (Tech.). Retrieved from http://www.cse.ucla.edu/products/reports/tr307.pdf

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

Silva Mangiante, E. M. (2011). Teachers matter: Measures of teacher effectiveness in low-income minority schools. *Educational Assessment, Evaluation and Accountability, 23*(1), 41-63. Retrieved from http://eric.ed.gov/?id=EJ913564

Smith, T. M., & Ingersoll, R. M. (2004). What are the effects of induction and mentoring on beginning teacher turnover. *American Educational Research Journal, 41*(3), 681-714. Retrieved from http://www.gse.upenn.edu/pdf/rmi/Effects-of-Induction-and-Mentoring-RMI-Fall-2004.pdf

Smith, G., & Schall, T. (2000). Do baseball players regress toward the mean? *The American Statistician, 54*, 231-245. Retrieved from http://www.jstor.org/stable/2685772?seq=1#page_scan_tab_contents

Soar, R. S., Medley, D. M., & Coker, H. (1983). Teacher evaluation: A critique of currently used methods. *Phi Delta Kappan,* 239-246.

Soy, S. K. (1997, Spring). The case study as a research method. Retrieved from https://www.ischool.utexas.edu/~ssoy/usesusers/l391d1b.htm

Spillane, J. P. (2006). *Distributed leadership*. San Francisco: Jossey-Bass.

State Collaborative on Reforming Education. (2012, June). *TVAAS: An introduction to value-added in Tennessee* (Rep.). Retrieved from http://tnscore.org/research-reports/archived-reports/

Statistical Analytical System (SAS). (2007). *Resource guide for value-added reporting* (Rep.). Retrieved from http://www.dpi.state.nc.us/docs/evaas/guide/resourceguide.pdf

Stronge & Associates. (2013, March). *Stronge teacher evaluation system: A validation report* (Rep.). Retrieved from http://www.cesa6.org/effectiveness_project/Validation-Report-of-Stronge-Evaluation-System.pdf

Stuit, D., Berends, M., Austin, M. J., & Gerdeman, R. D. (2014, January). *Comparing estimates of teacher value-added based on criterion- and norm-referenced tests* (Rep.). Retrieved from http://ies.ed.gov/ncee/edlabs/regions/midwest/pdf/REL_2014004.pdf

Sturman, M. C., Cheramie, R. A., & Cashen, L. H. (2005). The impact of job complexity and performance measurement on the temporal consistency, stability, and test-retest reliability of employee job performance ratings. *Cornell University School of Hotel Administration the Scholarly Commons*. Retrieved from http://scholarship.sha.cornell.edu/cgi/viewcontent.cgi?article=1117&context=articles

Taylor, E. S., & Tyler, J. H. (2011, March). *The effect of evaluation on performance: Evidence from longitudinal student achievement data of mid-career teachers* (Working paper). Retrieved from http://www.nber.org/papers/w16877.pdf

Taylor, E. S., & Tyler, J. H. (2012, Fall). *Can teacher evaluation improve teaching?* (Rep.). Retrieved from http://educationnext.org/files/ednext_20124_taylortyler.pdf

Teacher Leadership Exploratory Consortium. (2011). *Teacher leader model standards* (Rep.). Retrieved from https://www.ets.org/s/education_topics/teaching_quality/pdf/teacher_leader_model_standards.pdf

Tennessee Department of Education. (2012, July). *Teacher evaluation in Tennessee: A report on year 1 implementation* (Rep.). Retrieved from https://www2.ed.gov/programs/racetothetop/communities/tle2-year-1-evaluation-report.pdf

Tennessee Department of Education. (2014). First to the top. Retrieved from https://www.tn.gov/education/topic/first-to-the-top

Thompson, M. L. (1962). A historical approach to teacher evaluation. *The Clearing House, 37*(3), 169-172.

Turner, R. R. (1987). What teachers think about their evaluations. *Education Digest, 52*(6), 40-43. Retrieved from http://connection.ebscohost.com/c/articles/22020516/what-teachers-think-about-their-evaluations

Turner, D. (2010, April 8). States push to pay teachers based on performance. Retrieved from http://usatoday30.usatoday.com/news/education/2010-04-08-teachers-pay_n.htm

Turque, B. (2010, July 24). Rhee dismisses 241 D.C. teachers; union vows to contest firings. Retrieved from http://www.washingtonpost.com/wp-dyn/content/article/2010/07/23/AR2010072303093.html

U.S. Department of Education. (n.d.a). Goals 2000: Educate America Act. Retrieved from http://www2.ed.gov/legislation/GOALS2000/TheAct/index.html

U.S. Department of Education. (n.d.b). Section VII-Teacher Evaluation and Development. Retrieved from http://www.ed.gov/teaching/national-conversation/vision/section-vii-teacher-evaluation-and-development

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

U.S. Department of Education. (2010a). *A blueprint for reform the Reauthorization of the Elementary and Secondary Education Act*. Washington, DC: U.S. Department of Education, Office of Planning, Evaluation and Policy Development.

U.S. Department of Education. (2010b, December). The No Child Left Behind Act of 2001. Retrieved from http://www2.ed.gov/policy/elsec/leg/esea02/index.html

U.S. Department of Education. (2011a). *Bringing flexibility & focus to education law: Supporting state and local progress* (Rep.). Retrieved from http://files.eric.ed.gov/fulltext/ED527485.pdf (ERIC Document Reproduction Service)

U.S. Department of Education. (2011b, December 23). Department of Education awards $200 million to seven states to advance K-12 reform. Retrieved from http://www.ed.gov/news/press-releases/department-education-awards-200-million-seven-states-advance-k-12-reform

U.S. Department of Education. (2011c, September 23). Obama administration sets high bar for flexibility from no child left behind in order to advance equity and support reform. Retrieved from http://www.ed.gov/news/press-releases/obama-administration-sets-high-bar-flexibility-no-child-left-behind-order-advanc

U.S. Department of Education. (2014, March). Race to the Top fund. Retrieved from http://www2.ed.gov/programs/racetothetop/index.html

University of North Carolina. (2012, Fall). *Annual report on teacher education* (Rep.). Retrieved from http://www.ncleg.net/documentsites/committees/JLEOC/Reports%20Received/Archives/2012%20Reports%20Received/Teacher%20Education%20Report.pdf

Usher, A. (2011, December). *AYP results for 2010-11* [PDF]. Washington, DC: Center on Educational Policy.

Vescio, V., Ross, D., & Adams, A. (2008). A review of research on the impact of professional learning communities on teaching practice and student learning. *Teaching and Teacher Education, 24*, 80-91. Retrieved from http://www.k12.wa.us/Compensation/pubdocs/Vescio2008PLC-paper.pdf

Wake County Public School System. (2009, March). *Comparison of SAS© EVAAS© results and WCPSS effectiveness index results* (Rep.). Retrieved from https://webarchive.wcpss.net/results/reports/2009/0911evaas_index.pdf

Wei, H., Hembry, T., Murphy, D. L., & McBride, Y. (2012, May). *Value-Added models in the evaluation of teacher effectiveness: A comparison of models and outcomes* (Rep.). Retrieved from http://images.pearsonassessments.com/images/tmrs/ComparisonofValue-AddedModelsandOutcomes.pdf

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. (2009). *The widget effect* (Rep.). Retrieved from http://widgeteffect.org/

Wenglinsky, H. (2000). *How teaching matters: Bringing the classroom back into discussions of teacher quality*. Princeton, NJ: Policy Information Center, Educational Testing Service.

Winerip, M. (2011, November 6). In Tennessee, following the rules for evaluations off a cliff. *New York Times*. Retrieved from http://www.nytimes.com/2011/11/07/education/tennessees-rules-on-teacher-evaluations-bring-frustration.html?_r=0

Wise, A. E., Darling-Hammond, L., McLaughlin, M. W., & Bernstein, H. T. (1984). *Case studies for teacher evaluation: A study of effective practices*. Santa Monica, CA: Rand.

Wood, C. J., & Pohland, P. A. (1979). Teacher evaluation: The myth and realities. In W. D. (Ed.), *Planning for the evaluation of teaching* (pp. 73-82). Bloomington, IN: Phi Delta Kappa.

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

Wright, P. S., White, J. T., Sanders, W. L., & Rivers, J. C. (2010, March 25). *SAS EVAAS statistical models* (Rep.). Retrieved from http://www.sas.com/resources/asset/SAS-EVAAS-Statistical-Models.pdf

Xu, D. (2000). *The relationship of school spending and student academic achievement when achievement is measured by value-added scores* (Doctoral dissertation, Vanderbilt University, 2000) (pp. 1-129). Ann Arbor, MI: Bell & Howell Information and Learning Company.

Yin, R. K. (2009). *Case study research: Design and methods* (4th ed.). Thousand Oaks, CA: Sage Publications.

Yin, R. K. (2014) *Case study research: Design and methods* (5th ed.). Thousand Oaks, CA: Sage Publications.

Appendix A

NCEES Teacher Survey

IC1.
Informed Consent Form
Introduction

This survey will collect information about how the North Carolina Educator Evaluation System (NCEES) has affected your teaching practices and teacher leadership.

Procedures

The survey consists of some initial demographic questions and 18 survey questions, including 18 free-response questions that follow each survey question. The survey will take approximately 20 to 30 minutes, depending upon the number of free-response questions completed. This survey will be conducted using Qualtrics.

Risks/Discomforts

Risks are minimal to nonexistent for involvement in this study. Your participation is voluntary. You will not be asked for any personal information. Your name will not be connected to any response in this survey process. Your responses will be tracked using a numerical identifier generated and applied by Qualtrics.

Benefits

There are no direct benefits for participants. However, through your participation, researchers will learn more about the effects of the NCEES on teaching practices and teacher leadership.

Confidentiality

All data obtained from participants will be kept confidential. Your responses to the survey will mostly be reported in an aggregate format (by reporting combined results). Responses to free-response items on this survey may be quoted in part or in entirety; however, your responses to free-response items will be anonymous. All responses to this survey will be concealed, and no one other than the primary investigator and professor listed below will have access to them. The data collected will be stored in the HIPAA-compliant, Qualtrics-secure database until it has been deleted by the primary investigator.

Participation

Participation in this research study is voluntary. You have the right to withdraw at anytime or refuse to participate entirely without penalty. If you desire to withdraw, please close your Internet browser or window and your responses will not be recorded. Responses will be recorded only when the respondent reaches the end of the survey questions. It is not required that every survey and free-response question be answered. Survey or free-response questions not completed by respondents will simply be registered as "not completed" and will not harm the quality of responses to other items.

Questions about the Research

If you have questions regarding this study, you may contact Daniel Wydo at dawydo@clevelandcountyschools.org.

Questions about your Rights as Research Participants

If you have questions you do not feel comfortable asking the researcher, you may contact Dr. Doug Eury at aeury@gardner-webb.edu or 704-406-4402. You may also contact the director of Gardner-Webb University's Institutional Review Board at inaydenova@gardner-webb.edu or (704) 406-4242.

Q1.
I have read and understood the consent form and desire of my own free will to participate in this study.

- Yes
- No

Terms.

The following terms will be used in this survey and will reappear for review as a heading over each new block of questions:

NCEES - The North Carolina Educator Evaluation System (NCEES) is the recently implemented teacher evaluation system that formally replaced the revised Teacher Performance Appraisal Instrument (TPAI-R) during the 2011-2012 school year.

Teaching Practices - The teaching practices of teachers refer to the instructional techniques in your classroom with your assigned students (i.e., lesson planning, assessment strategies).

Teacher Leadership - Teacher leadership involves any assigned or volunteered activities outside your classroom that benefit other teachers and non-assigned students (i.e., level of participation in PLC, contribution to positive working conditions).

NCEES Standards:
Standard I: Teachers Demonstrate Leadership
Standard II: Teachers Establish a Respectful Environment for a Diverse Population of Students
Standard III: Teachers Know the Content They Teach
Standard IV: Teachers Facilitate Learning for Their Students
Standard V: Teachers Reflect on Their Practice
Standard VI: Teachers Contribute to the Academic Success of Their Students (calculated using EVAAS®).

Q2. Please provide your years of teaching experience (private/public - including any other county, state, or country)

1 - 5	6 - 10	11 - 15	16 - 20	21 - 25	26 - 30	More than 30
●	●	●	●	●	●	●

Q3. Please provide your gender.

Male	Female
●	●

Q4. Please provide your ethnicity/race.

Hispanic or Latino	Black	White	Native American or American Indian	Asian or Pacific Islander
●	●	●	●	●

Q5. I have received classroom observations and evaluation using a teacher evaluation system prior to the NCEES (e.g., under former NC systems: TPAI [1990's], TPAI-R [2000's], or under any other system from other states or countries).

Yes	No
●	●

Q6. I am a classroom teacher at the following school:

Q7. In using the NCEES, I have received an adequate number of formal classroom observations to rate my performance. (formal>20 minutes)

Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
○	○	○	○	○

Q8. How has this affected your teaching practices and teacher leadership?

Q9. In using the NCEES, I have received an adequate number of informal classroom observations (e.g., walk-throughs, snapshots) to rate my performance. (informal<20 minutes)

Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
○	○	○	○	○

Q10. How has this affected your teaching practices and teacher leadership?

Q11. In using the NCEES, there has been adequate time to participate in pre and post conferences.

Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
○	○	○	○	○

Q12. How has this affected your teaching practices and teacher leadership?

Q13. In using the NCEES, my ratings for Standards 1 - 5 align overall with my rating for Standard 6 (from EVAAS®).

Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree	I do not receive EVAAS scores or have not received enough data to form an opinion.
○	○	○	○	○	○

Q14. How has this affected your teaching practices and teacher leadership?

Q15. In using the NCEES, the rubric for Standards 2 - 5 has served as an appropriate guide to rate my *teaching practices*.

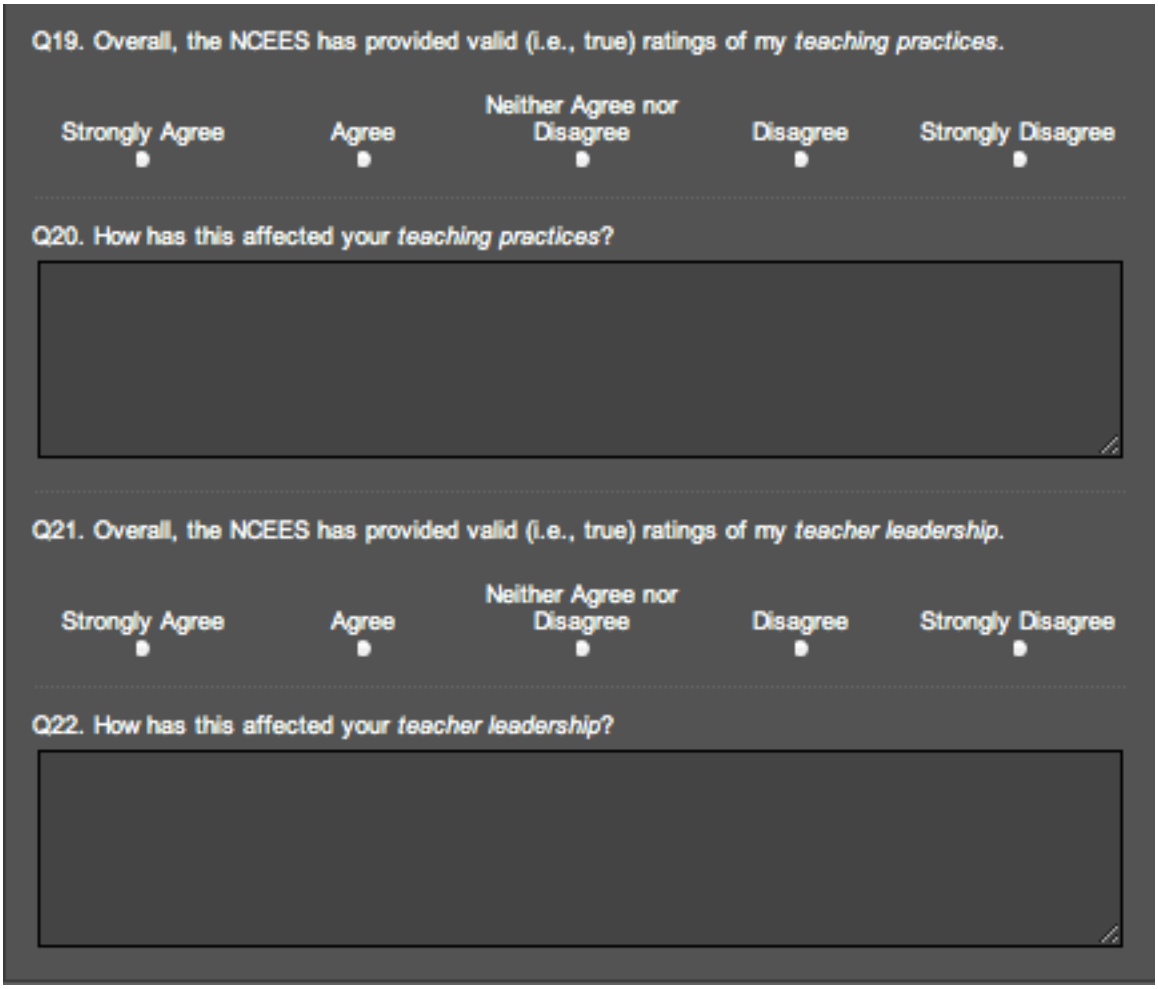
Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
○	○	○	○	○

Q16. How has this affected your *teaching practices*?

Q17. In using the NCEES, the rubric for Standard 1 has served as an appropriate guide to rate my *teacher leadership*.

Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
○	○	○	○	○

Q18. How has this affected your *teacher leadership*?


Q19. Overall, the NCEES has provided valid (i.e., true) ratings of my *teaching practices*.
Strongly Agree
Agree
Neither Agree nor Disagree
Disagree
Strongly Disagree
Q20. How has this affected your *teaching practices*?
Q21. Overall, the NCEES has provided valid (i.e., true) ratings of my *teacher leadership*.
Strongly Agree
Agree
Neither Agree nor Disagree
Disagree
Strongly Disagree
Q22. How has this affected your *teacher leadership*?

Q23. In using the NCEES, I have received reliable classroom observation ratings, consistent with my performance, for Standards 1 - 5 from semester to semester and year to year.

Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree	I am a newer teacher and have not received enough ratings to form an opinion.
○	○	○	○	○	○

Q24. How has this affected your teaching practices and teacher leadership?

Q25. In using the NCEES, various observers (different administrators/peers) have shown agreement in rating my performance for Standards 1 - 5 from semester to semester and year to year.

Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree	I am a newer teacher and have not received enough ratings to form an opinion.
○	○	○	○	○	○

Q26. How has this affected your teaching practices and teacher leadership?

Q27. In using the NCEES, Standard 6 (from EVAAS®) has provided a rating that reflected the academic growth of my students.

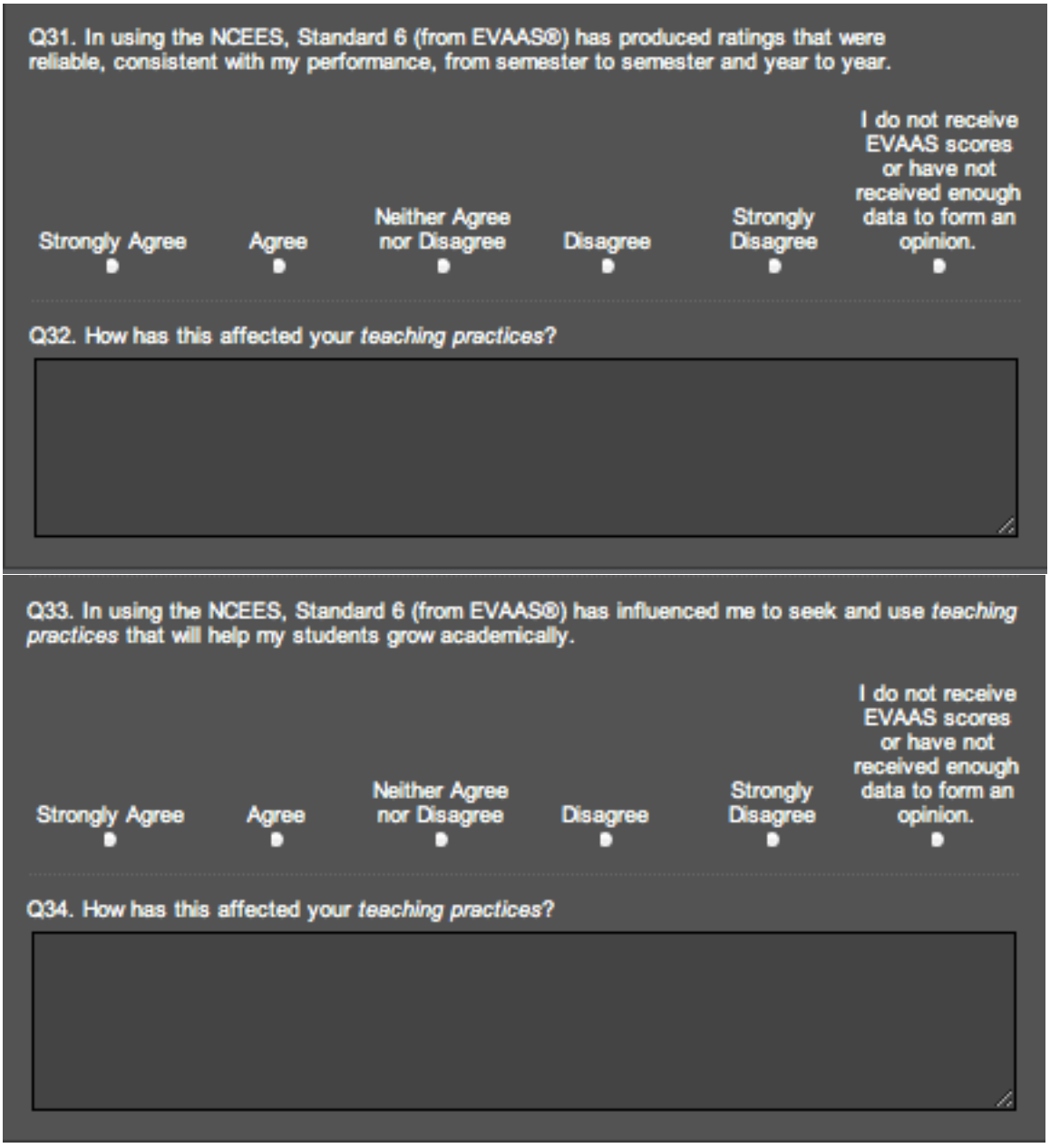
Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree	I do not receive EVAAS scores or have not received enough data to form an opinion.
○	○	○	○	○	○

Q28. How has this affected your teaching practices and teacher leadership?

Q29. In using the NCEES, Standard 6 (from EVAAS®) has provided a rating that reflected the quality of my *teaching practices*.

Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree	I do not receive EVAAS scores or have not received enough data to form an opinion.
○	○	○	○	○	○

Q30. How has this affected your *teaching practices*?



Q31. In using the NCEES, Standard 6 (from EVAAS®) has produced ratings that were reliable, consistent with my performance, from semester to semester and year to year.

Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree	I do not receive EVAAS scores or have not received enough data to form an opinion.
○	○	○	○	○	○

Q32. How has this affected your *teaching practices*?

Q33. In using the NCEES, Standard 6 (from EVAAS®) has influenced me to seek and use *teaching practices* that will help my students grow academically.

Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree	I do not receive EVAAS scores or have not received enough data to form an opinion.
○	○	○	○	○	○

Q34. How has this affected your *teaching practices*?

Q35. In using the NCEES, I have received useful feedback from observers to improve my *teaching practices*.

Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
○	○	○	○	○

Q36. How has this affected your *teaching practices*?

Q37. In using the NCEES, I have received useful feedback from observers to improve my *teacher leadership*.

Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
○	○	○	○	○

Q38. How has this affected your *teacher leadership*?

Q39. In using the NCEES, my summative ratings (at the end of the year) have helped to improve my *teaching practices*.

Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
○	○	○	○	○

Q40. How has this affected your *teaching practices*?

Q41. In using the NCEES, my summative ratings (at the end of the year) have helped to improve my *teacher leadership*.

Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
○	○	○	○	○

Q42. How has this affected your *teacher leadership*?

Appendix B

Survey Invitation for NCEES Teacher Survey

Dear Teacher,

You are invited to participate in a research study conducted by Daniel Wydo under the direction of Gardner-Webb University. The research project explores the effects of the North Carolina Educator Evaluation System (NCEES) on teaching practices and teacher leadership in the six high schools in this county. My hope is that you will be able and willing to add to the current research body aimed to improve teaching practices and teacher leadership.

The online survey from Qualtrics will take a minimum of twenty minutes and consists of twenty-two Likert-scale items. Every item gives you the ability to share personalized comments in an area underneath. You may share your comments if you have any. If you do not have any comments, please leave this area blank. Leaving any of the comment fields blank will not invalidate your survey–your "Agree"/"Disagree" choices on the scale are valuable and will still be recorded for analysis.

The survey time window opens now, at receipt of this email. This time window will remain open for two weeks and will be closed on ----------.

LINK:----------

This study has no foreseen risks apparent. You may refuse to participate. You may exit the survey at any time without penalty and your responses will not be recorded. There is no penalty for not participating or exiting out of the survey early. In order for the survey to be completed, you must pass through all twenty-two questions, regardless of whether any questions or comment fields are left blank–whichever questions were answered will be recorded once you have complied with the instructions at the end and agree to exit the finished survey.

Your participation in this survey is completely voluntary and confidential. I will not receive any identifying data connected with the survey results. This survey results will be accessible exclusively to select Gardner-Webb professors and myself. No identifying data will be available to Gardner-Webb professors. All aspects of your participation will be anonymous.

If you have any questions or technical concerns, please contact the researcher at dawydo@clevelandcountyschools.org. If you have any questions or concerns about the methods of research used in this survey, you may call the Gardner-Webb University Institutional Review Board at 704-406-4000.

Respectfully,

Daniel Wydo (dawydo@clevelandcountyschools.org)

Appendix C

North Carolina Educator Evaluation System (NCEES)
Principal Interview

1. Please provide the following background information:

    a. What is the approximate size of your school, in terms of numbers of students, teachers, and administrators?
    b. How many years were you a teacher and an administrator?
    c. How many different teacher evaluation systems did you operate under as a teacher and an administrator?

2. Please comment about how teacher performance is observed and evaluated in your school.

3. Please comment on the ability of the NCEES to measure teaching practices and teacher leadership in your school.

4. Please comment on how EVAAS® outcomes are used in your school.

5. Please comment on how outcomes from the NCEES are used in your school.

Appendix D

Goodlad's Eight Concluding Points
(Goodlad & Klein, 1970)

"One conclusion stands out clearly: many of the changes we have believed to be taking place in schooling have not been getting into classrooms; changes widely recommended for the schools over the past 15 years were blunted on school and classroom door.

Second, schools and classrooms were marked by a sameness regardless of location, student enrollment, and "typing" as provided initially to us by an administrator.

Third, there seemed to be a considerable discrepancy between teachers' perceptions of their own innovative behavior and the perceptions of observers. The teachers sincerely thought they were individualizing instruction, encouraging inductive learning, involving children in group processes, and so on.

Fourth, "special," supplementary, and enrichment activities and practices differed very little from "regular" classroom activities.

Fifth, general or specific classroom goals were not identifiable to observers. Instruction was general in character and not specifically directed to diagnosed needs, progress, and problems of individual children. Teachers shot with a shotgun, not a rifle.

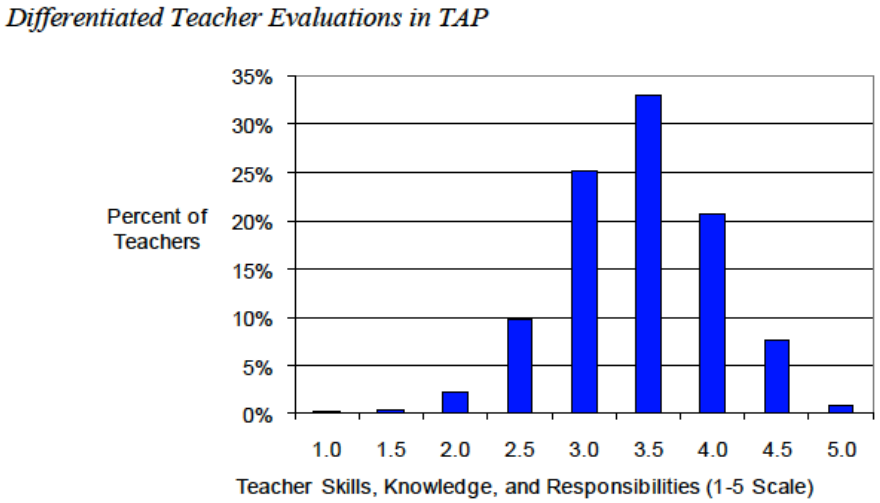
Sixth, the direction being pursued by the school as a whole was equally obscure or diffused.

Seventh, there appeared not to be a critical mass of teachers, parents, and others working together toward developing either a sense of direction or solutions to school-wide problems concerning them.

Eight parallels number seven: school personnel appeared to be very much alone in their endeavors. Principals tended to remain in offices and hallways and not to intrude on sacred classroom ground in any direct way. Teachers, although alone and presumably free to teach in their classrooms, appeared to be bound to a common conception of what school is and should be." (pp. 97-98)

Appendix E

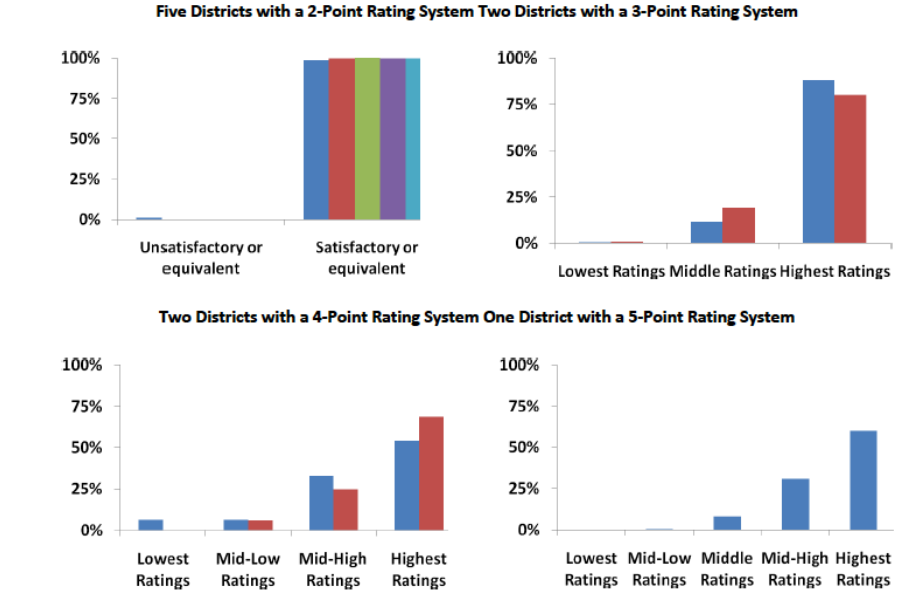
Distribution of Teachers Among Standards in TAP (Daley & Kim, 2010, p. 24)



Appendix F

Distribution of Teacher Performance:
Various Point Rating Systems from Weisberg et al. (2009) (Daley & Kim, 2010, p. 7)

Figure 1.Teacher Evaluation Ratings in Urban School Districts with 2-3-4-and 5-Point Rating Systems



Data from Weisberg et al., (2009).

Appendix G

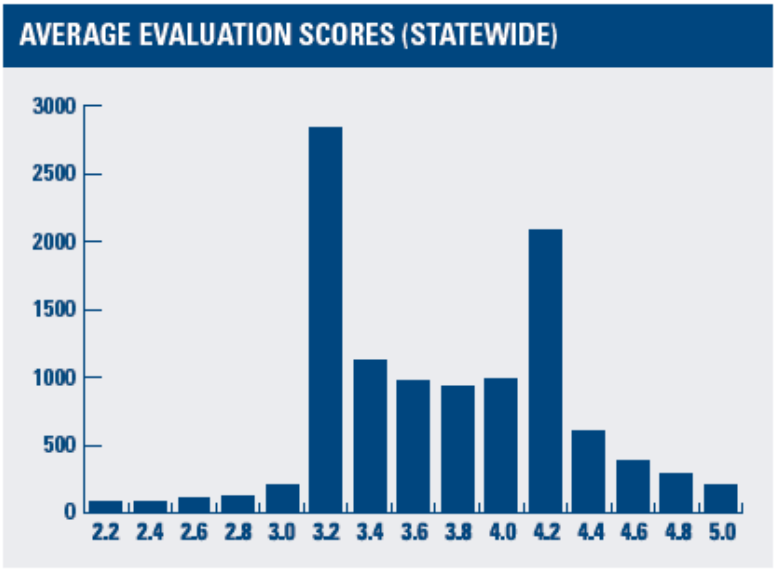
Distribution of Teacher Performance Under IMPACT Compared to Previous Year (Curtis, 2011, p.23)

Comparison of Teacher Ratings Under 1st Year of IMPACT and Previous Year Results

SY 2009-10 TEACHER IMPACT RATINGS	INEFFECTIVE	MINIMALLY EFFECTIVE	EFFECTIVE	HIGHLY EFFECTIVE
Distribution of Teacher Ratings SY 2009-10	3%	16%	66%	15%
SY 2008-09 TEACHER EVALUATION RATINGS	**UNSATISFACTORY**	**NEEDS IMPROVEMENT**	**MEETS EXPECTATIONS**	**EXCEEDS EXPECTATIONS**
Distribution of Teacher Ratings SY 2008-09	0.2%	4.8%	50%	45%

Appendix H

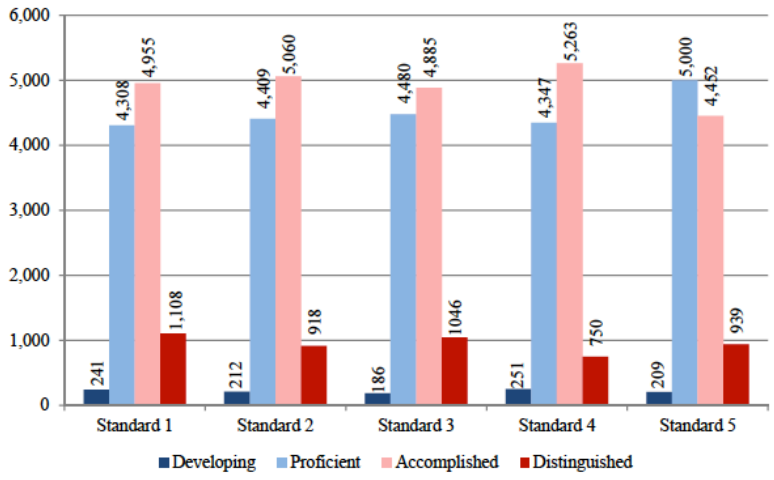
Distribution of Teacher Performance Under the NCEES:
Statewide Average Scores Across All Standards (Batten et al., 2012, p. 4)



Appendix I

Teacher Ratings For NCEES: Broken Down By Standard (Lynn et al., 2013, p. 24)

Figure 1. Number of Teachers in Each Rating Category by Standard[31]



Appendix J

Distribution of TVAAS Versus Principal Ratings
(TNDOE, 2012, p. 32)

Distribution of TVAAS Individual Teacher Effect and Observation Scores

Level	1	2	3	4	5
TVAAS Individual Teacher Effect	**16.5%**	**8.1%**	**24.5%**	**11.9%**	**39.1%**
Observation	**0.2%**	**2.2%**	**21.5%**	**53.0%**	**23.2%**

**Figures rounded to the nearest tenth.*

Appendix K

Table 1. EVAAS® Outcomes Versus Teacher Ratings on Standard 4 At Regional Level (Batten et al., 2012, p. 5)

Region	Coefficient	R^2
1	0.068	7.5%
2	0.058	5.5%
5	0.047	4.5%
7	0.046	3.2%
8	0.046	3.2%
4	0.042	4.5%
3	0.037	2.5%
6	0.036	2.3%

Appendix L

Table 2. EVAAS® Outcomes Versus Mean Teacher Ratings Across NCEES At Regional Level
(Batten et al., 2012, p. 5)

Region	Coefficient	R^2
1	0.067	8.9%
2	0.064	8.5%
8	0.057	6.0%
5	0.050	6.4%
7	0.048	4.3%
4	0.045	6.5%
3	0.040	3.8%
6	0.040	3.7%

Appendix M

Observation Ratings From Principals vs. Value-Added Measures
(Sartain et al., 2011, p. 56)

There is a significant relationship between observation ratings and value-added measures

Reading (N=795 observations of 417 teachers)					
Framework Component	Unsatisfactory	Basic	Proficient	Distinguished	Omnibus F-statistic
2a	-0.041	-0.162	0.226	0.264	5.33***
2b	-0.470	-0.086	0.186	0.411	6.60***
2c	-0.212	-0.083	0.170	0.364	4.67***
2d	-0.158	-0.04	0.175	0.326	2.94**
2e	-0.353	-0.026	0.180	0.430	4.99***
3a	-0.376	-0.054	0.191	0.305	4.13***
3b	-0.352	0.142	0.145	0.320	2.40*
3c	-0.111	-0.052	0.190	0.323	3.75**
3d	-0.338	0.005	0.239	0.391	5.39***
3e	-0.397	-0.087	0.201	0.429	6.87***

Mathematics (N=653 observations of 340 teachers)					
Framework Component	Unsatisfactory	Basic	Proficient	Distinguished	Omnibus F-statistic
2a	-0.030	-0.237	0.042	0.327	4.73***
2b	-0.552	-0.301	0.083	0.368	6.83***
2c	-0.105	-0.196	0.000	0.434	7.00***
2d	-0.359	-0.351	0.068	0.436	9.83***
2e	-0.165	-0.221	0.07	0.450	5.98***
3a	-0.639	-0.141	-0.011	0.370	6.60***
3b	-0.597	-0.043	0.085	0.299	3.09**
3c	-0.422	-0.160	0.062	0.335	4.10***
3d	-0.424	-0.076	0.08	0.424	4.48***
3e	-0.281	-0.104	-0.006	0.522	5.90***

Note: The table shows each of the ten components of teaching practice for which principals assigned ratings. The numbers in the rating columns are the average value-added measure for the teachers who received that rating in that component. For example in Component 3e, teachers with an Unsatisfactory rating had an average value-added measure of -0.397, which is more than one-third of a standard deviation below a teacher whose students achieved average student growth. The average teacher who received a Basic in 3e had a -0.087 value-added measure, Proficient had 0.201, and Distinguished had 0.429. In reading, for all components except 2a, the average value-added measure increases across the rating categories. The consistent correlation between the value-added measure and the classroom observation measure suggests that the Framework ratings are a valid measure of teacher practice.

Appendix N

NCEES Ratings vs. Teacher Characteristics, Including EVAAS (Lynn et al., 2013, p. 23)

Table 3. Teacher Characteristics by NC TEP Ratings[30]

	Not Demonstrated	Developing	Proficient	Accomplished	Distinguished
Standard 1	4	241	4,308	4,955	1,108
White	75.0%	71.5%	78.2%	81.0%	85.1%
Male	25.0%	29.5%	30.3%	24.5%	18.4%
Avg. Experience	5.50	6.98	8.59	11.74	14.92
EVAAS	-2.94	-2.06	-0.53	0.31	0.82
Standard 2	17	212	4,409	5,060	918
White	88.2%	78.0%	77.9%	81.2%	84.8%
Male	35.3%	30.6%	30.9%	23.9%	17.2%
Avg. Experience	17.71	8.11	8.76	11.61	15.16
EVAAS	-0.10	-2.28	-0.50	0.35	0.67
Standard 3	19	186	4,480	4,885	1,046
White	89.5%	76.7%	77.2%	81.4%	86.5%
Male	26.3%	29.3%	29.3%	24.7%	20.2%
Avg. Experience	19.84	6.72	8.40	11.89	15.58
EVAAS	0.53	-1.99	-0.56	0.33	0.92
Standard 4	5	251	4,347	5,263	750
White	50.0%	74.6%	76.1%	82.6%	86.8%
Male	25.0%	31.1%	30.6%	23.7%	18.5%
Avg. Experience	9.40	7.41	8.83	11.64	15.94
EVAAS	-3.80	-1.98	-0.61	0.41	0.87
Standard 5	16	209	5,000	4,452	939
White	100.0%	73.5%	78.9%	80.5%	85.7%
Male	31.3%	37.0%	30.3%	23.1%	18.0%
Avg. Experience	17.06	6.35	9.04	11.89	14.42
EVAAS	0.48	-1.86	-0.43	0.35	0.71

[30] Evaluation standards defined as: 1) Teachers demonstrate leadership, 2) Teachers establish a respectful environment for a diverse population of students, 3) Teachers know the content they teach, 4) Teachers facilitate learning for their students, and 5) Teachers reflect on their practice.

Appendix O

Round 1 Results–Correlation of Danielson's *Framework* (1996; 2007) With Student Outcomes at Various Test Sites (Milanowski et al., 2004, p. 4)

Round 1:

Average Correlations Between Teacher Evaluation Scores and Estimates of Achievement Within Classrooms For Three Research Sites

	Tested Subject		
Site	Reading	Math	Other
Cincinnati	.32	.43	.27 (Science)
Vaughn	.50	.21	.18 (Language Arts)
Washoe	.21	.19	-

Appendix P

Round 2 Results–Correlation of Danielson's *Framework* (1996; 2007) With Student Outcomes at Various Test Sites (Milanowski et al., 2004, pp. 13 – 14)

Round 2:

Correlations Between Teacher Evaluation Scores and Estimates of Average Student Achievement Based on Empirical Bayes Intercept Residuals from Models with Controls for Student Characteristics

	Tested Subject		
Site	Reading	Math	Other
Cincinnati	.28	.34	-0.2* (Science)
Vaughn	.61	.45	.38 (Language Arts)
Washoe	.25	.24	-

*Confidence Interval Includes 0

Appendix Q

Effect Size of an Increase in Teacher Rating for One Level of Improvement

(Milanowski et al., 2004, p. 14)

Effect of a One Level Change in Teacher Evaluation Score on Student Achievement, (in Standard Deviation Units)

	Tested Subject		
Site	Reading	Math	Other
Cincinnati	.14	.18	-.01 (Science)
Vaughn	.25	.37	.21 (Language Arts)
Washoe	.14	.19	-

Appendix R

Correlation of Danielson's *Framework* (1996; 2007) With Student Outcomes at Various Test Sites Including a Three Year Rolling Average (Heneman et al., 2006, p. 5)

Exhibit IV. Average Correlations Between Teacher Evaluation Ratings and Estimates of Average Student Achievement in Reading and Mathematics

		Subject	
Site	Grades		
Cincinnati		Reading	Math
2001-2002	3-8	.48	.41
2002-2003	3-8	.28	.34
2003-2004	3-8	.29	.22
3-year average:		.35	.32
Coventry			
1999-2000	2,3,6	.17	.01
2000-2001	2,3,4,6	.24	-.20
2001-2002	4	.29	.51
3-year average:		.23	.11
Vaughn			
2000-2001	2-5	.48	.20
2001-2002	2-5	.58	.42
2002-2003	2-5	.05	.17
3-year average:		.37	.26
Washoe			
2001-2002	3-5	.21	.19
2002-2003	4-6	.25	.24
2003-2004	3-6	.19	.21
3-year average:		.22	.21

[2] Our research is reported in several journal articles and a book chapter. Validity results can be found in Kimball, White, Milanowski, and Borman (2004); Milanowski (2004); Gallagher (2004); and Milanowski, Kimball, and Odden (2005). Research on the implementation of the systems and teacher reactions can be found in Heneman and Milanowski (2003); Kimball (2002); and Milanowski and Heneman (2001).

Appendix S

Relationship Between Student Test Scores & Teacher TES
(Kane et al., 2010, p. 41)

Table 4: Estimates of the Relationship Between Student Test Scores & Teacher

TES Score Principal Components

	(A) Math					
	(1)	(2)	(3)	(4)	(5)	(6)
1. Overall Classroom Practices	0.543**	0.221**	0.202**	0.105**	0.275**	0.362**
	(0.108)	(0.041)	(0.037)	(0.032)	(0.037)	(0.117)
2. Classroom Environment Relative to	0.231+	0.128*	0.121*	0.082*	-0.021	0.023
Instructional Practices	(0.122)	(0.051)	(0.051)	(0.040)	(0.088)	(0.329)
3. Questions & Discussion Approach	0.065	0.001	-0.009	0.001	-0.031	0.051
Relative to Standards & Content Focus	(0.140)	(0.060)	(0.061)	(0.039)	(0.097)	(0.219)
Student Controls		Y	Y	Y	Y	Y
Teacher Experience Controls			Y	Y	Y	Y
School Fixed Effects				Y		
Teacher Fixed Effects						Y
Teacher Sample	207	207	207	207	49	49
Student Sample	16,196	16,196	16,196	16,196	4,109	4,109
Adjusted R-squared	0.049	0.527	0.529	0.556	0.545	0.561
	(B) Reading					
	(1)	(2)	(3)	(4)	(5)	(6)
1. Overall Classroom Practices	0.487**	0.209**	0.214**	0.141**	0.106+	0.294*
	(0.094)	(0.036)	(0.037)	(0.029)	(0.055)	(0.134)
2. Classroom Environment Relative to	0.160	0.059	0.059	0.036	0.201+	0.212
Instructional Practices	(0.145)	(0.049)	(0.049)	(0.039)	(0.106)	(0.157)
3. Questions & Discussion Approach	0.184	0.109*	0.100*	0.061	0.011	-0.297
Relative to Standards & Content Focus	(0.135)	(0.046)	(0.046)	(0.038)	(0.072)	(0.236)
Student Controls		Y	Y	Y	Y	Y
Teacher Experience Controls			Y	Y	Y	Y
School Fixed Effects				Y		
Teacher Fixed Effects						Y
Teacher Sample	365	365	365	365	81	81
Student Sample	20,125	20,125	20,125	20,125	5,251	5,251
Adjusted R-squared	0.039	0.528	0.529	0.550	0.550	0.582

Note: Each column represents a separate student-level specification. Clustered (teacher)

standard errors in parentheses. **p<0.01, *p<0.05, +p<0.1.

Appendix T

Stronge et al.'s (2013) Correlation Measurements: Standards 1-6 and 7 (p. 10)

Table 6. Correlations Between the Six Process Standards and the Achievement Standard

Standard	**Standard 1:** Professional Knowledge	**Standard 2:** Instructional Planning	**Standard 3:** Instructional Delivery	**Standard 4:** Assessment of and For Student Learning	**Standard 5:** Learning Environment	**Standard 6:** Professionalism
Standard 7: Student Academic Progress	.361**	.417**	.326**	.380**	.266**	.378**

Appendix U

Stronge et al.'s (2013) Correlation Measurements: Standards 1-6 (p. 10)

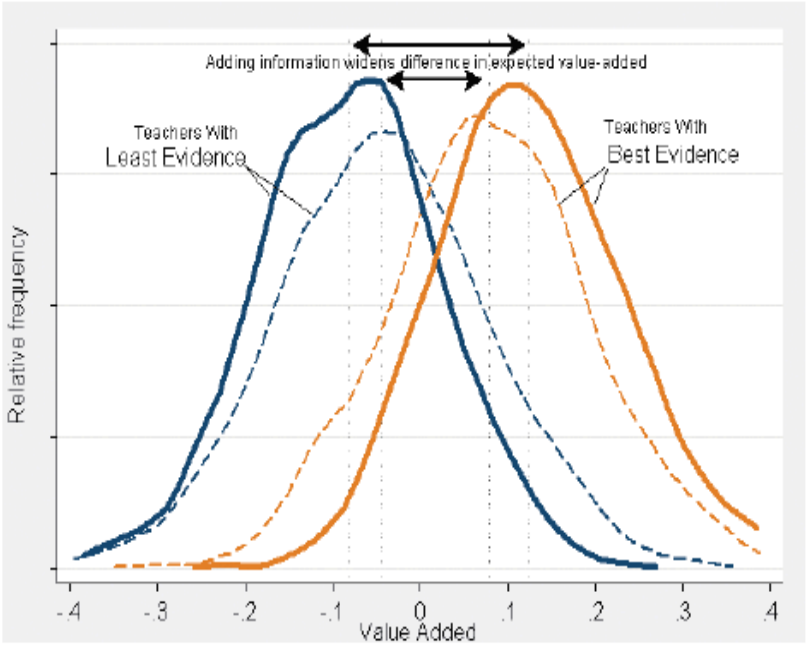
Table 5. Correlations Among the Six Stronge Teacher Process Standards

Standard	**Standard 1:** Professional Knowledge	**Standard 2:** Instructional Planning	**Standard 3:** Instructional Delivery	**Standard 4:** Assessment of and for Student Learning	**Standard 5:** Learning Environment	**Standard 6:** Professionalism
Standard 1: Professional Knowledge	1					
Standard 2: Instructional Planning	.585**	1				
Standard 3: Instructional Delivery	.628**	.535**	1			
Standard 4: Assessment of and for Student Learning	.617**	.623**	.576**	1		
Standard 5: Learning Environment	.563**	.535**	.547**	.519**	1	
Standard 6: Professionalism	.658**	.531**	.517**	.552**	.472**	1

Appendix V

MET (2010) Graph:
Showing Widened Distribution of Teacher Performance Based on Student Survey Data and VAM (p. 30)

Figure 1. Adding Value-Added Widens the Difference in Expected Effectiveness and Reduces Overlap



Note: The figure reports the distribution of value-added scores for the quarter of teachers with the most and least evidence of effectiveness. The dotted lines refer to the distributions based only on student perceptions. The solid lines report the distributions when value-added data from another section are added.

Appendix W

Lynn et al.'s (2013) Regression Estimates: Standards 1-5 Compared To Teacher Ratings, Experience, and Other Controls (p. 26)

Table 5: EVAAS and NC TEP Standards, OLS Estimates

	Standard 1	Standard 2	Standard 3	Standard 4	Standard 5
Rating	Coefficient (Std. Err.)	Coefficient (Std. Err.)	Coefficient (Std. Err.)	Coefficient (Std. Err.)	Coefficient (Std. Err.)
Not Demonstrated	-1.07 (1.32)	**2.04*** (0..66)	**2.37*** (0.63)	-1.71 (1.17)	**2.02*** (0.68)
Proficient	**1.41*** (0.17)	**1.65*** (0.18)	**1.29*** (0.20)	**1.27*** (0.17)	**1.20*** (0.19)
Accomplished	**2.18*** (0.18)	**2.42*** (0.19)	**2.12*** (0.20)	**2.21*** (0.17)	**1.90*** (0.19)
Distinguished	**2.67*** (0.19)	**2.71*** (0.20)	**2.71*** (0.21)	**2.66*** (0.19)	**2.23**** (0.21)
Teacher Experience					
< 3years	-0.09 (0.09)	-0.15 (0.09)	-0.09 (0.09)	-0.08 (0.09)	**-0.19*** (0.09)
3-5 years	0.08 (0.08)	0.07 (0.08)	0.09 (0.08)	0.09 (0.08)	0.07 (0.08)
11-25 years	**-0.18*** (-0.07)	**-0.16*** (-0.07)	**-0.21**** (0.07)	**-0.18*** (0.07)	**-0.17*** (0.07)
>25 years	-0.16 (0.11)	-0.11 (0.11)	-0.21 (0.11)	-0.13 (0.11)	-0.12 (0.11)
Controls					
Male	**-0.43*** (0.06)	**-0.44*** (0.06)	**-0.45*** (0.06)	**-0.42*** (0.06)	**-0.43*** (0.06)
Black	**-0.48*** (0.07)	**-0.52*** (0.07)	**-0.47*** (0.07)	**-0.43*** (0.07)	**-0.52*** (0.07)
Other Race	-0.05 (0.13)	-0.01 (-0.13)	-0.02 (0.13)	0.00 (0.13)	-0.07 (0.13)
Constant	**-1.60*** (0.18)	**-1.81*** (0.19)	**-1.51*** (0.20)	**-1.56*** (0.18)	**-1.28*** (0.20)
n	10,616	10,616	10,616	10,616	10,616
R^2	0.06	0.06	0.06	0.07	0.05

Note: Reference groups are: ***NC TEP Indicator***—developing rating level; ***Teacher Experience***— 6-10 years' experience; ***Controls***— female and White.

Note: * indicates statistically significant at the p<0.05 level; ** indicates statistically significant differences at the p<0.01 level; *** indicates statistically significant differences at the p<0.001 level.

Appendix X

Inter-rater Reliability from Stronge & Associates (2013) (p. 19)

Table 13. Individual Ratings Compared to Calibrated Ratings

	Rating at Calibrated Level	Rating within One Level of Calibration Level	Rating Two Levels away from Calibrated Level
Standard 1	196 (74.5%)	67 (25.5%)	0
Standard 2	213 (81.6%)	48 (18.4%)	0
Standard 3	236 (87.4%)	34 (12.6%)	0
Standard 4	234 (88.0%)	32 (12.0%)	0
Standard 5	197 (74.6%)	67 (25.4%)	0
Standard 6	222 (82.8%)	46 (17.2%)	0
Standard 7	261 (87.3%)	38 (12.7%)	0

Appendix Y

Intra-reliability Measurements from Strong et al. (2013) (p. 21)

Table 15: Reliability Coefficient of Performance Ratings in Each School

School	Reliability Coefficient						
A	0.92	G	0.89	M	0.79	T	0.99
B	0.98	H	0.91	N	0.78	U	0.82
C	0.87	I	0.78	O	0.92	V	0.79
D	0.76	J	0.94	P	0.84	W	0.80
E	0.87	K	0.92	Q	0.90	X	0.83
F	0.90	L	0.42	R	0.75	Y	0.66

Appendix Z

Ratings of Principals Compared to External Observers from Sartain et al. (2011, p. 14)

TABLE 3

Distribution of ratings for principals and observers

Rating	Principal (N=4,747 ratings)	Observer (N=4,852 ratings)
Distinguished	803 (17%)	157 (3%)
Proficient	2,530 (53%)	3,259 (67%)
Basic	1,291 (27%)	1,343 (28%)
Unsatisfactory	123 (3%)	93 (2%)

Appendix A1

Percent Improvement in Reliability When Adding Observers (Ho & Kane, 2013, p. 15)

Table 6

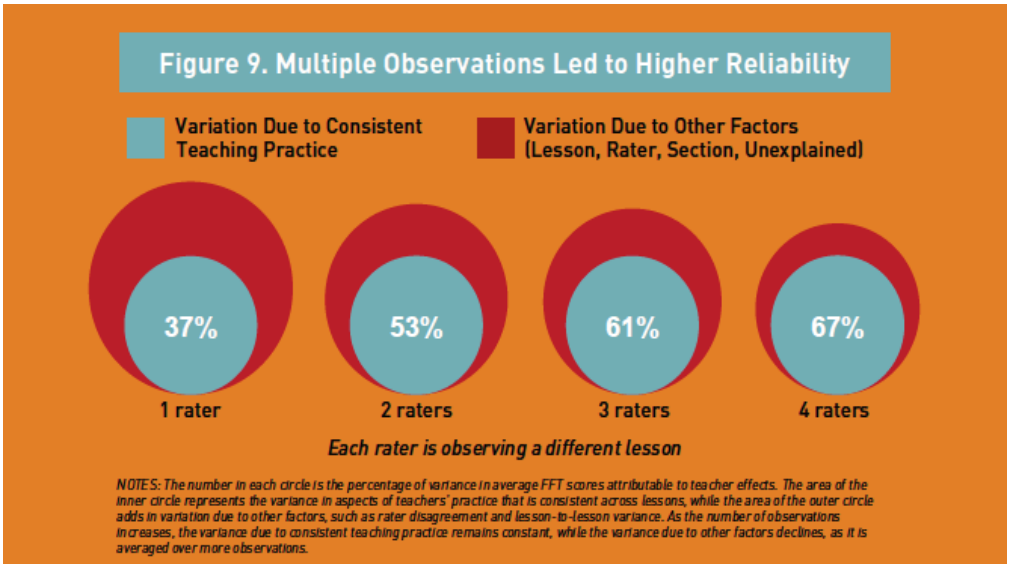
MEAN SCORES, SOURCES OF VARIANCE, AND RELIABILITY BY TYPE OF OBSERVER

			FROM G-STUDY			PERCENT IMPROVEMENT IN RELIABILITY		
BY TYPE OF RATER	Mean Score	# of Ratings	S.D. of Teacher Effect	SEM of a Single Observation	Reliability of 1 Rating by 1 Observer	1→2 Lessons (1 rater)	1→2 Raters (1 lesson)	1→2 Raters (2 lessons)
Administrator from same school	2.6658	428	0.326	0.320	0.510	14.22%	22.72%	31.56%
Difference relative to own administrator:								
Administrator from another school	-0.10	844	0.332	0.380	0.433	17.45%	35.73%	39.13%
Peers from same grade range	-0.20	684	0.195	0.308	0.287	19.36%	42.15%	54.59%
Peers from different grade range	-0.25	606	0.190	0.338	0.240	21.96%	47.22%	60.65%

Notes: Sample was limited to chosen lessons, along with a set of four lessons from the teachers who were indifferent to lesson selection. Based on average scores for items 2a through 3e of the Framework for Teaching.

Appendix B1

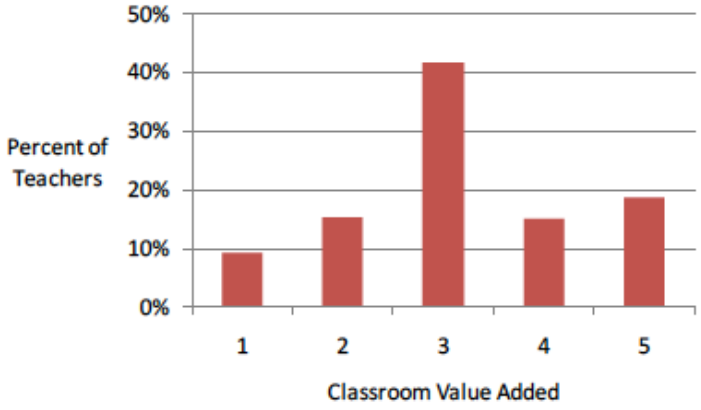
Resulting Reliability When Adding Multiple Observers (MET, 2012, p. 37)



Appendix C1

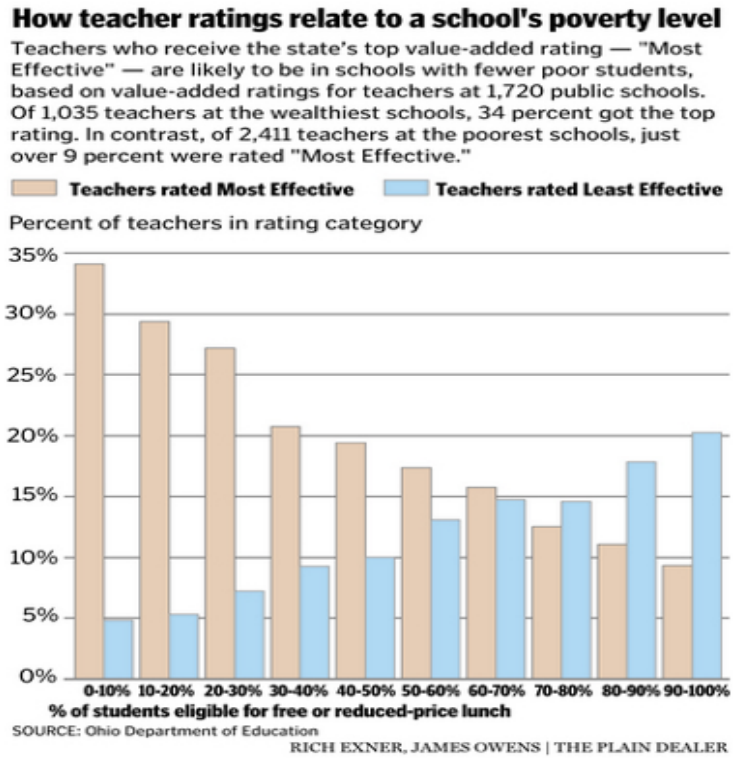
Distribution of Teacher Performance Judged By EVAAS® (Daley & Kim, 2010, p. 26)

Figure 6:
Distribution of Classroom Value Added for TAP Teachers in States Using EVAAS Scores, 2008



Appendix D1

O'Donnell's (2013) EVAAS® Teacher Ratings Versus School Poverty Level



Appendix E1

Correlation Measurements for the WCPSS Index Versus EVAAS®, the WCPSS Index Including Free/Reduced Lunch Rates As Covariate, and EVAAS® Including Free/Reduced Lunch Rates As Covariate (WCPSS, 2009, p. 4)

Table 1: Correlations between School-Level WCPSS Effectiveness Indices, EVAAS© School Value Added Estimates, and Percentage of Students in the School Eligible for Free or Reduced Price Lunch (FRL), 2007-08 School Year.

Test	WCPSS-EVAAS©	WCPSS-FRL	EVAAS©-FRL
Grade 4 Math (n=96)	0.97*	0.03	0.05
Grade 7 Math (n=25)	0.77*	0.02	-0.57*
Algebra I (n=24)	0.63*	0.03	-0.74*
Algebra II (n=23)	0.83*	0.03	-0.40*
Biology (n=21)	0.90*	-0.36	-0.64*
Chemistry (n=20)	0.97*	-0.01	-0.02
Civics & Econ. (n=20)	0.36	0.14	-0.64*
English I (n=23)	0.95*	-0.28	-0.46*
Geometry (n=23)	0.99*	-0.29	-0.38
Physical Science (n=20)	0.79*	0.00	-0.54*
Physics (n=14)	0.92*	-0.05	-0.40
US History (n=23)	0.81*	-0.18	-0.40*
All Tests Combined (n=332)	**0.79***	**-0.05**	**-0.30***

NOTE: Correlations with an asterisk (*) imply that there is evidence of a relationship between the two variables. Sample sizes for correlations are indicated in parentheses after each test name. Algebra I data are from high schools only.

Appendix F1

Nature of Disagreement Between EVAAS® and the WCPSS Index (WCPSS, 2009, p. 6)

Table 3: Nature of Classification Disagreements between EVAAS© School Value-Added reports and WCPSS Effectiveness Index, 2007-08 School Year.

WCPSS	EVAAS©	n	% of Total Comparisons
Low	Average	3	0.9%
Low	High	1	0.3%
Average	Low	31	9.3%
Average	High	91	27.4%
High	Low	0	-----
High	Average	6	1.8%
No disagreement		**200**	**60.2%**

NOTE: Results based only on the 12 tests listed in Table 1 (332 total school comparisons). Percentages do not add up to exactly 100% due to rounding.

Appendix G1

Schools Where EVAAS® and WCPSS Index Were More Favorable (WCPSS, 2009, p. 7)

Table 4: Classification Disagreements between EVAAS© School Value-Added reports and WCPSS Effectiveness Index by School Poverty Level, 2007-08 School Year.

Nature of Classification Disagreement	Total # Instances	# Instances Where School >= 30% FRL
EVAAS© more favorable	95	21 (22%)
No Disagreement	200	94 (47%)
WCPSS more favorable	37	25 (68%)

NOTE: "FRL" = Free or Reduced-Price Lunch.

Appendix H1

Various VAM Models Used in Goldhaber and Theobald (2013, p. 3) Analysis

Table 1: Large Vendors that Estimate Teacher Effectiveness Using Student Test Scores

Vendor	Name of Model	Brief Description
American Institutes for Research (AIR)	Varied	In most situations, models control for student background
Mathematica	Varied	In most situations, models control for student background
National Center for the Improvement of Educational Assessment (NCIEA)	Student Growth Percentile (SGP) Models	Models a descriptive measure of student growth within a teacher's classroom
SAS	EVAAS	Models control for prior test scores but not other student background variables
Value Added Research Center (VARC)	Varied	In most situations, models control for student background

Appendix I1

Outcomes For Each VAM model in Goldhaber and Theobald (2013, p. 7) Analysis

Table 3: Average Percentile Rankings in Advantaged and Disadvantaged Classrooms

Panel 1: Math	**Advantaged**	**Disadvantaged**
Student Growth Percentiles	60.7	41.1
VAM with prior test score	65.1	38.2
VAM with prior test score and student covariates	57.8	47.7
VAM with prior test score, student, and classroom covariates	60.1	46.6
VAM with within-school comparison	51.9	48.7
Panel 2: Reading	**Advantaged**	**Disadvantaged**
Student Growth Percentiles	66.6	33.8
VAM with prior test score	71.8	29.0
VAM with prior test score and student covariates	58.2	43.6
VAM with prior test score, student, and classroom covariates	60.3	42.8
VAM with within-school comparison	51.0	49.4

Appendix J1

Lauen et al.'s (2013) Regression Outcomes for Student and School Characteristics Versus VAM Rankings for Teachers (p. 17)

Table 1. Teacher Value-Added Index, by Student, Classroom, and School Characteristics

	Mathematics			
Model:	(1)	(2)	(3)	(4)
Minority	-0.040 **	-0.035 *	-0.039 **	-0.035 *
Poor	-0.075 ***	-0.053 ***	-0.073 ***	-0.053 ***
Prior Achievement	0.099 ***	-0.011 ***	0.098 ***	-0.011 ***
Class Avg. Prior Achievement		0.425 ***		0.425 ***
School Pct. Poor			-0.211 ***	0.002
Observations	363,824	363,824	363,824	363,824
Variance Components				
Within Schools	5.213	5.146	5.213	5.146
Between Schools	3.160	3.160	3.161	3.161
Percent Between-School Variance	37.7%	38%	37.8%	38.1%

	Reading			
Model:	(1)	(2)	(3)	(4)
Minority	-0.012 *	-0.008	-0.011	-0.007
Poor	-0.037 ***	-0.028 ***	-0.036 ***	-0.027 ***
Prior Achievement	0.028 ***	-0.005 ***	0.027 ***	-0.005 ***
Class Avg. Prior Achievement		0.143 ***		0.142 ***
School Pct. Poor			-0.182 ***	-0.104 ***
Observations	421,498	421,498	363,824	363,824
Variance Components				
Within Schools	1.162	1.153	1.162	1.153
Between Schools	0.806	0.806	0.806	0.806
Percent Between-School Variance	40%	41.1%	41%	41.1%

Note: * denotes sig diff at $p<.05$, ** at $p<.01$, *** at $p<.001$.

Appendix K1

Newton et al.'s (2010) Correlation Results Comparing Teacher VAM Rankings and Student Characteristics (p. 11)

Table 5

Significant Correlations between Teachers' VAM Rankings and Their Students' Characteristics, 2007

	ELL		Meal		Asian		Hispanic		Parent Ed.	
	Math	ELA	Math	ELA	Math	ELA	Math	ELA	Math	ELA
M1	-.38***	-.48***	-.27*	-.45***	.27*	.31***	-.33**	-.43***	.34**	.48***
M2	-.37***	-.31***	-.25*	-.20*	.18	.24*	-.27*	-.26**	.28**	.32**
M3	-.37***	-.42***	-.30**	-.30**	.31**	.30**	-.35**	-.39***	.35**	.38***
M4	-.31**	-.31**	-.31**	-.18	.24*	.31**	-.32**	-.30**	.32**	.31**
M5	-.29**	-.36***	-.34**	-.22*	.29**	.29**	-.34**	-.34***	.34**	.32***

*$p < .10$. **$p < .05$. ***$p < .01$.

Appendix L1

Newton et al.'s (2010) Subset of Teachers and VAM Outcomes With High Tracked Versus Untracked Students in Deciles (p. 14)

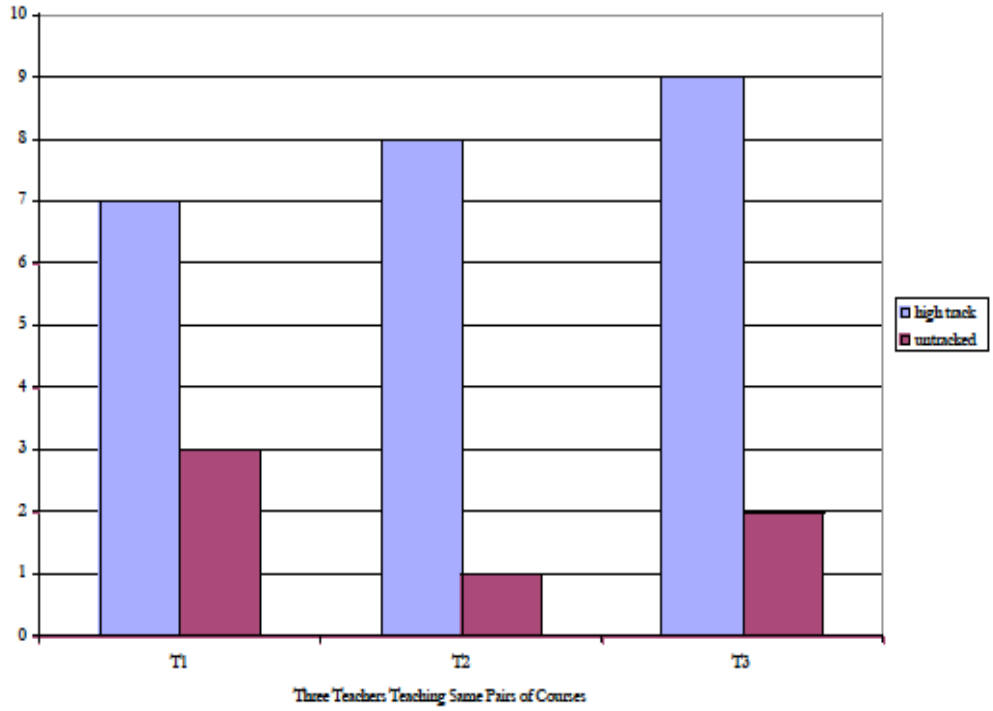


Figure 1: Teachers Rankings by Courses

Appendix M1

Ballou's (2005) Reliability of VAM Rankings Among Quartiles–1998 Versus 1999 (p. 17)

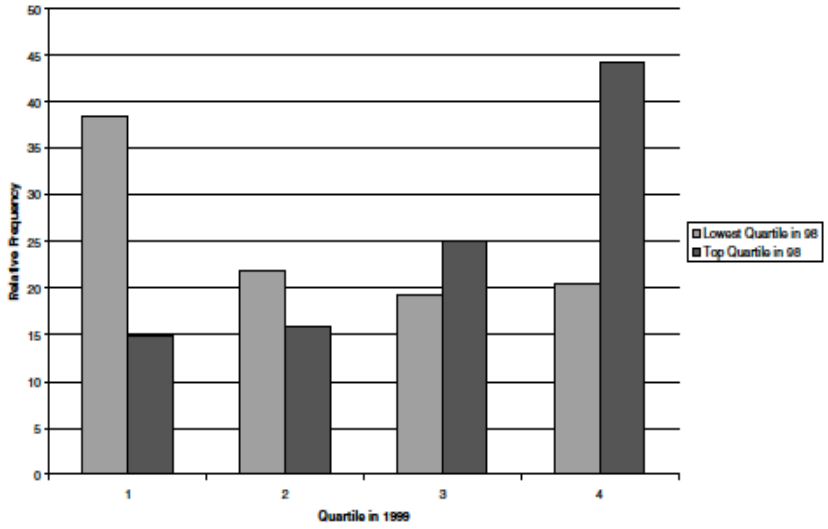


Figure 5a. Stability of Teacher Effects: Reading

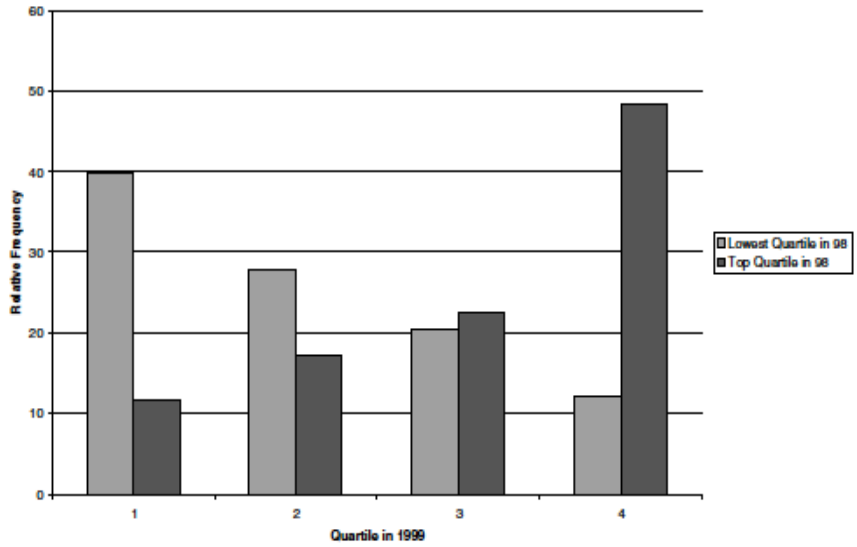


Figure 5b. Stability of Teacher Effects: Math

Appendix N1

Koedel and Betts (2007) Reliability Measures Across Quintiles Over Two Years (p. 30)

Table 7. Persistence of Teacher Fixed Effects Estimates across Data Subsets (Percentages)

		Teacher Coefficient Quintile Ranking From Year t				
		1	2	3	4	5 (best)
Teacher Coefficient Quintile Ranking From Year t-1	1	**30**	20	19	18	13
	2	23	25	13	21	18
	3	18	20	25	24	13
	4	15	16	26	20	23
	5 (best)	13	17	16	19	**35**

Note: (N = 941). Teachers are placed into quintiles using coefficient estimates from each data subset separately, quintile 5 being the best. Rows sum to 100 percent.

Appendix O1

Aaronson et al. (2007) Reliability Measures Across Quartiles Over Two Years (p. 114)

Table 5
Quartile Rankings of Estimated Teacher Effects in Years *t* and *t* + 1: Percent of Teachers by Row

Quartile in year *t*:	Quartile in Year *t* + 1			
	1	2	3	4
1	36	29	26	10
2	24	31	32	12
3	20	32	23	24
4	8	12	23	57

NOTE.—χ^2 test of random quartile assignment: $p < .000$. Quartile rankings are based on teacher effects estimated for each year based on the specification in col. 1 of table 6.

Appendix P1

McCaffrey et al.'s (2009) Pooled Year-to-Year Pairwise Correlation of Teacher VAM Outcomes Based On Various Controls and Persistence for Five Counties (p. 589)

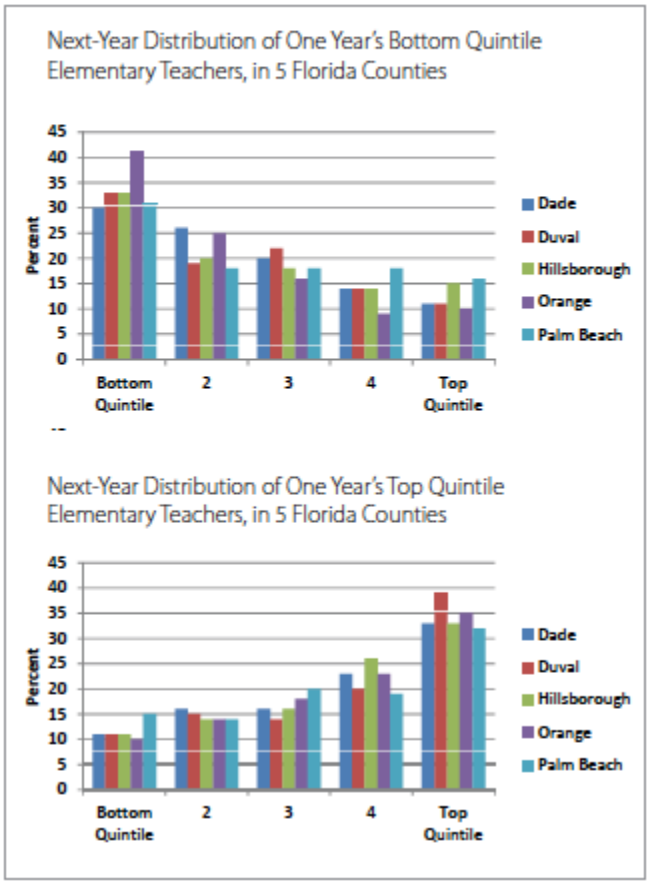
Table 2. Pooled Year-to-Year Pairwise Correlations of Estimated Teacher-by-Year Effects by County, Grade Level, and Model Type, 2000–1 to 2004–5 (Teachers with ≥15 Students in a Year)

County	Model Type: Student Controls	Model Type: Persistence	Pooled Year-to-Year Pairwise Correlations
		Elementary	
Dade	Covariates	Complete	.32
	Covariates	Partial	.46
	Fixed effects	Complete	.23
Duval	Covariates	Complete	.22
	Covariates	Partial	.30
	Fixed effects	Complete	.26
Hillsborough	Covariates	Complete	.27
	Covariates	Partial	.35
	Fixed effects	Complete	.24
Orange	Covariates	Complete	.34
	Covariates	Partial	.43
	Fixed effects	Complete	.39
Palm Beach	Covariates	Complete	.31
	Covariates	Partial	.44
	Fixed effects	Complete	.16
		Middle	
Dade	Covariates	Complete	.37
	Covariates	Partial	.67
	Fixed effects	Complete	.38
Duval	Covariates	Complete	.38
	Covariates	Partial	.53
	Fixed effects	Complete	.32
Hillsborough	Covariates	Complete	.32
	Covariates	Partial	.47
	Fixed effects	Complete	.30
Orange	Covariates	Complete	.32
	Covariates	Partial	.61
	Fixed effects	Complete	.33
Palm Beach	Covariates	Complete	.29
	Covariates	Partial	.51
	Fixed effects	Complete	.28

Appendix Q1

Haertel's (2013) Analysis of McCaffrey et al.'s (2009) Year-to-Year Changes in Teacher Rankings Based on VAM Outcomes (p. 18)

Figure 4
Year-to-Year Changes in Teacher Value-Added Rankings Reported by McCaffrey et al. (2009, Table 4, p. 591)



Appendix R1

Sass's (2008) Comparison of Teacher Rankings in the Same Year Using Two Different Standardized Tests (NRT Versus SSS) (p. 5)

TABLE 3. CROSS-EXAM STABILITY OF ELEMENTARY MATH TEACHER EFFECT ESTIMATES BY QUINTILE (PERCENT OF TEACHERS BY ROW) – HILLSBOROUGH CO., FL, 2001/02 [CORRELATION = 0.48]

	Ranking Based on SSS				
Ranking based on NRT	Bottom 20%	Second 20%	Third 20%	Fourth 20%	Top 20%
Bottom 20%	43	26	14	11	5
Second 20%	26	25	23	17	10
Third 20%	15	21	21	24	18
Fourth 20%	11	19	25	20	24
Top 20%	4	9	17	27	43

NRT is the Stanford Achievement test, SSS is the "Sunshine State Standards" criterion-based exam tied to Florida's curriculum standards.

Appendix S1

Reliability Measurements Using Correlation of Different Section and Prior Year Data on Various Standardized Tests (MET, 2010) (p. 18)

Table 5. The Stable Component in Value-Added on Various Assessments

		DIFFERENT SECTION			PRIOR YEAR	
VARIABLE	TOTAL VARIANCE ONE SECTION	CORRELATION COEFFICIENT	IMPLIED VARIANCE OF STABLE COMPONENT	TOTAL VARIANCE PRIOR YEAR	CORRELATION COEFFICIENT	IMPLIED VARIANCE OF STABLE COMPONENT
TYPE OF TEST	[S.D. IN BRACKETS]		[S.D. IN BRACKETS]	[S.D. IN BRACKETS]		[S.D. IN BRACKETS]
State Math Test	0.053	0.380	0.020	0.040	0.404	0.016
	[0.231]		[0.143]	[0.20]		[0.127]
State ELA Test	0.032	0.179	0.006	0.028	0.195	0.005
	[0.178]		[0.075]	[0.166]		[0.073]
BAM Test	0.071	0.227	0.016			
	[0.266]		[0.127]			
Stanford 9 OE ELA	0.129	0.348	0.045			
	[0.359]		[0.212]			

Note: The standard deviation (s.d.) in value-added is the square root of the variance. The BAM scores and Stanford 9 scores were not available for any teacher in the year prior to the study.

Appendix T1

Correlation Among Various Measures of Teacher Effectiveness (MET, 2010, p. 22)

Table 6. The Stability of Effectiveness Measures Between-Sections Taught by the Same Teacher, 2009–10

VARIABLE	CORRELATION [SAMPLE SIZE]	VARIABLE	CORRELATION [SAMPLE SIZE]
VALUE ADDED MEASURES:		TRIPOD:	
VA on State Math Test	0.381 *** [520]	Sum of 7 C's	0.668 *** [956]
VA on State ELA Test	0.180 *** [574]	Control+Challenge	0.601 *** [956]
VA on BAM Test	0.228 *** [452]	Other 5 C's	0.682 *** [956]
VA on Stanford 9 OE ELA	0.348 *** [514]	Care	0.669 *** [956]
		Control	0.657 *** [956]
		Clarify	0.557 *** [956]
		Challenge	0.642 *** [956]
		Captivate	0.685 *** [956]
		Confer	0.614 *** [956]
		Consolidate	0.648 *** [955]

Note: The sample size for each correlation is reported inside the square brackets. A *,**, or *** indicates a correlation that is significantly different from zero at the .10, .05 and .01 level respectively.

Appendix U1

Regression outcomes from Lynn et al. (2013) Showing EVAAS® Outcomes Versus Principal Ratings Referencing 11-25 Years of Experience and Controlling For Female and White Demographics (p. 25)

Table 4: EVAAS and NCTEP Rating (Accomplished or Distinguished Indicator), OLS Estimates

Variable	Coefficient (Std. Err.)
NCTEP Indicator	
Demonstrate Leadership	**0.26**** (0.08)
Respectful Environment	0.15 (0.08)
Content Knowledge	**0.32***** (0.08)
Facilitate Learning	**0.57***** (0.08)
Reflect on Practice	0.00 (0.08)
Teacher Experience	
< 3yrs	-0.03 (0.09)
3-5 years	0.12 (0.08)
11-25 years	**-0.18*** (0.07)
>25 years	-0.14 (0.11)
Controls	
Male	**-0.41***** (0.06)
Black	**-0.44***** (0.07)
Other Race	-0.02 (0.13)
Constant	**-0.51***** (0.07)
n	10,616
R^2	0.06

Note: Reference groups are: ***NC TEP Indicator***—developing rating level; ***Teacher Experience***— 6-10 years' experience; ***Controls***— female and White.

Note: * indicates statistically significant at the p<0.05 level; ** indicates statistically significant differences at the p<0.01 level; *** indicates statistically significant differences at the p<0.001 level.