ASSESSING THE INTER-RATER RELIABILITY OF A SYSTEM-WIDE TEACHER EVALUATION OBSERVATION INSTRUMENT: MOVING BEYOND THE KAPPA PARADOX

by
ALBERT MANUEL JIMENEZ
(Under the Direction of SALLY J. ZEPEDA)

ABSTRACT

The purpose of this study was to investigate the processes and procedures impacting the validation and the establishment of inter-rater reliability of observation instruments used in a teacher evaluation context. A newly-created observation instrument used in an extensive teacher evaluation program in a southeastern school district served as the object of investigation in this research. Inter-rater reliability coefficients of the instrument were assessed as part of the validation of the evaluation system.

Additionally, two methods of inter-rater reliability that correct for chance agreement were examined to determine if the Gwet AC1 statistic, which is often used in a medical context but rarely in an education one, outperformed the typically provided kappa statistic. The inter-rater reliability coefficients for all videos and all items combined were in an acceptable range. This was also the case for most individual standards as well. Gwet's AC1 statistic regularly outperformed the kappa statistic as a chance-corrected measure of inter-rater reliability. This finding held for all teachers combined, for the highest-rated teacher, and for the lowest-rated teacher, suggesting that

Gwet's AC1 statistic shows promise for future inter-rater reliability studies in a teacher evaluation context.

While Gwet's AC1 statistic outperformed kappa for the lowest-rated teacher, what was clear is the inter-rater reliability coefficients for the lowest-rated teacher suggests that consistently and accurately identifying poorly performing teachers is elusive. Additionally, this finding suggests the possibility that standards by which teachers are traditionally assessed enabling accurate identification for poor performing teachers may be underdeveloped. Further research in this area is warranted.

INDEX WORDS: Classroom Observation Instruments, Teacher Evaluation, Teacher Evaluation System Creation, Validity, Inter-rater Reliability, Gwet's AC1 Statistic & Teacher Evaluation

ASSESSING THE INTER-RATER RELIABILITY OF A SYSTEM-WIDE TEACHER EVALUATION OBSERVATION INSTRUMENT: MOVING BEYOND THE KAPPA PARADOX

by

ALBERT MANUEL JIMENEZ
BA, Augusta State University, 2000
MS, Mississippi State University, 2004

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

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA
2014

ASSESSING THE INTER-RATER RELIABILITY OF A SYSTEM-WIDE TEACHER EVALUATION OBSERVATION INSTRUMENT: MOVING BEYOND THE KAPPA PARADOX

by

ALBERT MANUEL JIMENEZ

Major Professor: SALLY J. ZEPEDA
Committee: ALLAN S. COHEN
NÖEL GREGG

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2014

DEDICATION

This dissertation is dedicated to all of the people in my personal life that have assisted and inspired me to complete this accomplishment. First, my wife, Anisa, has always supported me in every way to be better than I thought I could be. I could never have completed this dissertation without her. She has made many sacrifices to allow me the needed time to complete this degree and always has done so with a great deal of positive encouragement. In addition to my wife, Van and Ariana, my children, are truly inspirations and give me the desire to be an example to them. My family is everything to me.

In addition, others in my family deserve much credit. I would like to thank my parents, who raised me to value education and who taught me that bypassing college was never an option. My wife's parents, Thomas and Rani Sullivan, are also owed so much. I could not have asked for more supportive people in my life and the support and love they have given so freely has meant the world to me.

I am very blessed to have such wonderful people in my life.

ACKNOWLEDGEMENTS

The work completed in pursuit of my doctoral degree could not have been completed without the help of many. While not enjoying myself and fighting through a doctoral program, fate intervened and my path crossed with Dr. Sally Zepeda. Once we were working together again, the steps to graduating became very clear and this is due in large part to her. She has supported me, given me many opportunities to advance professionally, and been a friend and mentor anyone would be lucky to have.

I am also thankful for the guidance of my remaining committee members, Dr. Allan Cohen and Dr. Nöel Gregg. They assisted me in focusing my study and provided direction that really contributed to me being able to complete this degree in a reasonable amount of time.

I also thank Dr. Allan Cohen and Dr. Steve Cramer for providing me with the opportunity to work at a terrific research center while completing this degree. They have truly improved the life I am able to help provide for my family. This is a true gift and one that allowed me to focus on my studies, knowing that my family was well taken care of while I pursued my degree.

Finally, I would like to acknowledge and thank the Southern Regional Education Board for providing me with three years of study through their fellowship program. This was also an invaluable gift that allowed me to focus on my studies.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Broadly, this study is designed to contribute to the research and discussion about the processes and procedures impacting the overall validation and the establishment of inter-rater reliability of observation instruments used in a teacher evaluation context. This topic is timely given that teacher evaluation systems around the country are being revamped in response to conditions set forth in the 2009 federal grants program, Race to the Top (RTTT). RTTT, a segment of the American Recovery and Reinvestment Act of 2009, was designed, in part, to increase the effectiveness of teaching and educational leadership (Clifford & Ross, 2011; Lohman, 2010). As such, these newly revamped teacher evaluation systems, which undeniably contain new observation instruments, must be validated and an assessment of reliability must accompany the validation of any rating instrument.

While RTTT has played a large role in spurring the implementation of newly-created teacher evaluation systems, existing research has also been foundational to establishing the need for higher quality teacher evaluation systems (Graham, Milanowski, & Miller, 2012). Having quality teachers has been identified as perhaps the most important element in having students achieve at high levels (Aaronson, Barrow, & Sander 2007; Buddin & Zamarro, 2009; Darling-Hammond, 1997; Mendro, 1998; Rockoff, 2004; Sanders, Wright, & Horn, 1997; Stronge & Tucker, 2000; Tucker & Stronge, 2005; Wenglinsky, 2002). The value of quality teaching and the need to determine more

accurately what that means has been a mainstream topic of discussion since *A Nation at Risk*:

> salary, promotion, tenure, and retention decisions should be tied to an effective evaluation system that includes peer review so that superior teachers can be rewarded, average ones encouraged, and poor ones either improved or terminated. (National Commission on Excellence in Education, 1983, p. 30)

With this focus on the topic of quality teaching in RTTT, the need for quality evaluation systems has intensified.

The need for newly designed, more accurate methods to determine teacher quality is clear, because "as states, districts, and schools transition toward more rigorous educator evaluation systems, they are placing additional weight on judgments about educator practice" (Graham et al., 2012, p. 4). Any study geared at determining the quality of a rating instrument is charged with gathering and interpreting the validity evidence that is in line with the intended uses of the observation instrument under examination, as it is the uses of the scores from the instrument, and not the instrument itself, that is being validated (Kane, 2009, 2013; Messick, 1995).

While the validation of the uses of the instrument is an important process, what is clear is that the first step is to "state the claims that are being made (explicitly or implicitly) in a proposed interpretation or use" (Kane, 2013, p. 7). Cook and Beckman (2006) agreed and argued, "because the validity of an instrument's scores hinges on the construct, a clear definition of the intended construct is the first step in any validity evaluation" (p. 166.e8). As such, a brief discussion of the purposes and intents of teacher evaluation is warranted.

According to Haefele (1993), there are seven basic functions which any teacher evaluation system should serve. These seven basic functions are, (1) screen out

unqualified persons, (2) provide constructive feedback to individual teachers, (3) recognize and reinforce high-performing teaching, (4) provide direction to staff development programs, (5) provide strong evidence which can stand judicial scrutiny, (6) assist in terminating unproductive personnel, and (7) unify teachers and administrators in the quest to educate students in a highly effective way. What is clear from an examination of these purposes as defined by Haefele (1993) is that they can be fairly easily divided into two categories, namely a teacher evaluation system is meant to serve both a summative and formative function.

The summative and formative functions that teacher evaluation systems are meant to serve can further be defined in terms of the tasks associated with each, the first is quality assurance and the second, is professional development (Danielson & McGreal, 2000; Marzano, 2012). The quality assurance task is basically accountability-based to identify accurately the quality of each teacher and to assist in making personnel decisions. The professional development task is to identify more precisely the professional shortcomings of individual teachers and to assist them in addressing these shortcomings, and in turn, become more highly-functioning professional educators (Marzano, 2012; Mead, Rotherham, & Brown, 2012).

The quality assurance task associated with teacher evaluation systems has been defined as evaluation (Danielson & McGreal, 2000). This is the function of the evaluation process which often meets legislated accountability requirements (McColskey & Egelson, 1993; Mead et al., 2012) and is used to identify and to rate teacher quality. While the rating of teachers is a main function of the teacher evaluation process, it is important to note that the summative function is found to be problematic in the literature

(Danielson & McGreal, 2000; Loup, Garland, Ellett, & Rugutt, 1996; McGreal, 1983; Wise, Darling-Hammond, McLaughlin, & Bernstein, 1984; Zepeda, 2006, 2013).

The formative function of teacher evaluation systems, often referred to as supervision, is to support the professional development and growth of individual teachers within a school (Danielson & McGreal, 2000; Mead et al., 2012; Stronge & Tucker, 2000; Zepeda, 2012). This function is often overlooked because of a lack of resources and time (McColskey & Egelson, 1993). While research has shown the formative aspects of the supervision and evaluation process is what is often neglected. To reach the goal of having highly qualified teachers in every classroom will require a renewed focus on the relationship between teacher evaluation, professional development, and supervision offered in a coherent manner (Zepeda, 2012). As Zepeda (2006) pointed out, "the new mantra for supervision in this millennium and beyond should be we must do more to break the outmoded culture of neglect, where evaluation for accountability supplants supervision that promotes growth and development" (p. 68). The purposes and intents of teacher evaluation should lay the foundational work for any teacher evaluation system, and the focus should remain on both the summative and formative aspects of the process (Stronge & Tucker, 2000).

This discussion of the purposes and intents of teacher evaluation is illustrative of one overarching idea which is central to this study. The purpose of teacher evaluation is, according to existing literature, to assess more accurately teacher quality for continued employment decisions or to identify more accurately professional development needs of teachers being evaluated (Danielson & McGreal, 2000; Marzano, 2012). In either case, the key to successful teacher evaluation is the word, *accurately*.

It is not just the assessment, but rather the accurate assessment of teacher quality which enables the purposes of the teacher evaluation process to be achieved. The observation of teaching is the "mainstay of teacher evaluation systems" (Zepeda, 2013, p. 65); however, without valid instruments the process means very little. The determination of how accurate the instruments used in the observation process are at assessing teacher quality is, ultimately, the purpose of establishing the reliability and validity of these instruments so that the ratings assigned to teachers in such high-stakes situations as continued employment can be trusted. Establishing the validity of rating instruments is foundational to every other aspect of teacher evaluation.

It is important to note that the process of validating an instrument is a multistep process though need not be what Messick (1989) referred to as an "endless process" (p. 151). The concept of validity has evolved since the belief that validation was never-ending, primarily through the work of Kane (1992, 2006, 2013). The argument-based approach to validity, which includes the concept of the interpretive/use argument (the IUA), lays out the reasoning behind the proposed uses of an instrument has a well-defined process. As Kane (2013) stated, "first, state the claims that are being made (explicitly or implicitly) in a proposed interpretation or use (the IUA), and second, evaluate these claims (the validity argument)" (p. 7).

While there are several types of evidence supporting validity claims which may be gathered, the types relevant to individual studies are dictated by the type of instrument being reviewed (Cook & Beckman, 2006; Kane, 1992, 2006, 2013; Messick, 1989). In validating the uses of a rating instrument, gathering information about the content relevance and representativeness of the construct is pivotal (Messick, 1995) because one

main threat to establishing validity is construct underrepresentation (Downing & Haladyna, 2004; Messick, 1999). As such, in a teacher evaluation context, the content with which teachers are rated must be assessed to determine if the standards contained on the instrument accurately and exhaustively cover the components of quality teaching.

While an assessment of the degree to which the content is accurately covered is one of the types of related validity evidence relevant to a rating instrument, the other main types of evidence to be gathered include 1) an assessment of the reliability of the observation instrument, 2) an assessment of the quality control procedures related to scoring, and 3) an assessment of potential negative consequences which could result from the use of the instrument. The assessment of the quality control procedures and potential unacceptable negative consequences are primarily completed through an argument based on the quality of implemented procedures.

The reliability assessment is accomplished, for this type of instrument, by determining if rater training and the instrument are adequate so that those charged with rating educators are able to consistently and accurately determine the quality of those being rated while using the observation instrument (Cook & Beckman, 2006). Graham et al. (2012) have argued one way to gather data to complete the reliability assessment and to determine the degree to which raters are able to identify consistently and accurately quality teaching "is to have raters rate a common set of video clips or artifacts" (p. 13). The type of reliability evidence which is particularly relevant to a rating instrument is inter-rater reliability (Cook & Beckman, 2006; Graham et al., 2012).

Inter-rater reliability is a statistical technique which assesses the degree to which raters are consistent in assigning ratings. There are several ways to calculate inter-rater

reliability. For example, one can calculate the percent of agreement, which is the percentage of the time that raters exactly agreed with one another (Cook & Beckman, 2006; Graham et al., 2012). Other types of inter-rater reliability coefficients include, but are not limited to, Cohen's kappa (1968) (and variations of), Phi, intraclass correlation coefficients (ICC), and Gwet's AC1 statistic (Cook & Beckham, 2006; Graham et al., 2012; Gwet, 2002, 2012). Many of these statistics are situational and are determined by factors such as the number of raters (Graham et al., 2012).

Often, in an educational context, inter-rater reliability is assessed through one of the many variations of Cohen's kappa and by reporting the percent of overall exact agreement (Graham et al., 2012). While this is commonplace, there is some question about the effectiveness of the various kappa statistics. Kappa statistics suffer from a paradox where high values of overall agreement can produce low values of kappa (Gwet, 2012). A more detailed discussion of inter-rater reliability as well as the various coefficients can be found in the review of the relevant literature in Chapter 2.

In addition to understanding the purposes of teacher evaluation and how these purposes impact the design of evaluation instruments, it is also important to acknowledge the impact the political arena can have in this area. In fact, Stronge (2006) labeled politics as "another major obstacle to effective teacher evaluation" (p. 12). Currently, the trend, resulting in large part from RTTT, has been to judge teacher performance primarily on student performance on standardized exams through the use of value-added models (VAMs) (Newton, Darling-Hammond, Haertel, and Thomas, 2010). They (Newton et al., 2010) note, "in the current policy climate, pupil learning is increasingly conceptualized as standardized test score gains" (p. 3).

While VAMs are gaining in popularity in the political arena, they have often been found to be statistically problematic (Koedel & Betts, 2011; McCaffrey, Sass, Lockwood, & Mihaly, 2009; Newton et al., 2010; Papay, 2011; Sass, 2008). The issues with VAMs and the impact political pressure can have on the evaluation of teachers are discussed in Chapter 2.

The accountability movement in public education is intended to increase the quality of education by holding all involved to higher standards. Ultimately, improving the quality of education through an increased focus on teacher quality and evaluation is a worthwhile goal, but one that creates an increased need to focus on the validity and reliability of the uses of the scores gathered from the instruments used to assess teacher quality. This study is designed to contribute to the literature about establishing the validity and reliability of educational rating instruments by 1) outlining the types of evidence which contribute to establishing the validity of teacher rating instruments and 2) discussing the issues related to the reliability of teacher rating instruments including both factors which can contribute to unreliability and also by illustrating the problems and advantages with various inter-rater reliability coefficients.

Statement of the Problem

The requirement to improve the evaluation of teacher quality in RTTT has resulted in an increased need to create new evaluation instruments designed to better measure teacher quality. One major problem with teacher evaluation systems of the past has been a lack of consistency with ratings (Danielson, 2011). The lack of consistency has likely been because performance assessments in an educational context are often designed and implemented before methodological issues are examined and addressed

(Linn & Baker, 1996). As such, outlining clear methods for validating these instruments, including establishing the inter-rater reliability, are of increased importance. This is because "observation ratings inherently rely on evaluators' professional judgment" and "there is always a question of how much the ratings depend on the particular evaluator rather than the educator's actual performance" (Graham et al., 2012, p. 4). As is discussed, this concern can be addressed through assuring that ratings are consistent across raters.

The process of establishing validity through the collection of construct evidence and by calculating the relevant reliability coefficients should include gathering evidence aligned to the purposes of the construct (i.e., assessing teacher quality), gathering evidence supporting that the construct is adequately measured, and by assessing reliability through several quantitative methods. This is all necessary in order to determine whether or not the scores gathered from the system can be trusted in the situations for which they were created.

While state agencies may have the means and knowledge to conduct the validation process for teacher observation instruments, very often local education agencies (LEAs) lack the statistical and methodological expertise to complete such a study. Results of this study should address this need of LEAs by outlining the steps of the validation process for teacher evaluation observation instruments, and doing so in a way that is clearly definable and replicable.

A secondary issue addressed in this study is related to the often reported inter-rater reliability coefficients in an educational context. Much of the assessment of inter-rater reliability has been completed by reporting the percentage of overall agreement

among raters and some variation of Cohen's kappa. While Cohen's kappa is regularly reported in reliability studies, this statistic, as well as the variations created for myriad rater situations, has been found to suffer from a possible paradox which is artificially deflated (Feinstein & Cicchetti, 1990; Gwet, 2012). In fact, Gwet (2012) claimed "kappa often yields coefficients that are unexpectedly low if compared to the overall percent agreement" (p. 36).

The increased need to validate these newly created teacher evaluation instruments and the possible paradox which may occur in the different variations of the kappa statistic have created a need in an educational context to provide more stable statistics establishing inter-rater reliability. In this study, we will illustrate the issues with kappa statistics and provide an alternative method to establish inter-rater reliability in educational contexts (Feinstein & Cicchetti, 1990; Gwet, 2002, 2012).

Purpose of the Study

The purpose of this research is two-fold. The first is to describe validation and reliability procedures in terms of a teacher evaluation context so that observation instruments can be rigorously assessed for quality. The second purpose of this study was to determine if currently used measures of inter-rater reliability are adequate when assessing inter-rater reliability in an educational context. The typical measures of inter-rater reliability reported in educational research often include a variation of Cohen's kappa (1968) statistic and the percent of overall agreement. Further, an additional measure, Gwet's AC1 statistic (2002), which has been shown in other contexts to be a more stable measure of inter-rater reliability than Cohen's kappa (1968), will be evaluated to determine if this is the case in a teacher evaluation context.

In this research, the variation of Cohen's kappa (1968) used is Fleiss' kappa (1971). This is appropriate when there are more than two raters as is typically the case in a teacher evaluation context. The additional measure of inter-rater reliability investigated in this study is Gwet's AC1 statistic (2002). Gwet's AC1 is a more robust estimate of inter-rater reliability, providing similar information and interpretation to that of Fleiss' kappa (1971) without suffering from the similar paradox (Gwet, 2002). This statistic has gained in popularity, especially for use for ratings in the medical field. Its applicability in a teacher evaluation context is examined in this study.

Background of the Study

The current research study is motivated by the creation of a new teacher evaluation system used in the Clarke County School District (CCSD) in Athens, GA. The new teacher evaluation system was a collaborative effort between the CCSD and Professor Sally Zepeda, a teacher supervision and evaluation expert at the University of Georgia. This system is one that addresses the requirements of RTTT by evaluating teachers across many areas but also had the goal of ultimately including student achievement data or perhaps other forms of artifacts and evidence into the evaluation process. The teacher evaluation system had a scoring system and rubric; however, the teacher evaluation observation instrument needed to be validated through contextual and statistical evidence, and it needed to be assessed in terms of its effectiveness in identifying differences in teacher quality across four rating scores (Unsatisfactory, Emerging, Proficient, or Exemplary) as defined by the Clarke County School District.

Prior to undertaking the steps necessary to create the new evaluation system, the CCSD had used the state evaluation instrument, the Georgia Teacher Observation

Instrument (GTOI), when conducting teacher evaluations. This instrument conceived in 1984 can be described as "minimal" at best; however, this system was the state sanctioned system and assessed teachers in three areas—1) provides instruction, 2) assesses and encourages student progress, and 3) manages the learning environment. The rating in these areas was either Needs Improvement or Satisfactory.

While a needs improvement rating often resulted in a teacher being placed on a professional development plan, very little within the evaluation system offered any constructive feedback for those receiving satisfactory ratings. Moreover, the GTOI did little to promote the professional growth of those rated as satisfactory. The need for more detailed evaluations which provide more useful information to assist in the growth of all teachers was apparent and resulted in the development of the Clarke County School District Teacher Evaluation System.

In 2012, the Clarke County School District Teacher Evaluation System was implemented. The Clarke County School District Teacher Evaluation System is a model that situates the teacher as a life-long learner. Teachers are placed into one of two tiers, based on experience and performance. Tier I teachers include all teacher in years 1, 2, or 3, teachers new to the school district, and teachers placed on a formal plan of improvement. Tier II teachers include all teachers in years 4, 5, or 6 and experienced teachers new to the district (at the principal's discretion). For this study, only Tier I and Tier II classroom observation instruments were evaluated. Tier III, set to be implemented in the 2013-2014 school year, was not implemented because in spring, 2013, HB 244 was passed in which all school systems must adopt by fall 2014 the state sanctioned teacher

evaluation system, the Teacher Keys Effectiveness System (TKES) and for leaders, the Leader Key Effectiveness System (LKES).

As compared with the three standards previously assessed using the state evaluation instrument, teachers in the Clarke County School District are now rated, based on their classroom observations, on either 6 (Tier I) or 7 (Tier II) performance standards which have been agreed on by the system as covering the requirements of being a high quality, professional educator in a standards-based classroom. The standards for Tier I included *Curriculum and Planning (CP), Standards-Based Instruction (SBI), Assessment of Student Learning (ASL), Instruction Environment (IE), Building Positive Student Relationships (BPSR) and Artifacts and Evidence (AE)*. Tier II teachers are assessed on the same 6 standards as Tier I teachers, though there is an additional standard, *Teacher Leader (TL)*, which asserts that as a teacher becomes a more seasoned veteran educator, they have a responsibility to assume additional responsibilities benefitting the school community and/or their peers.

Within each of the performance standards, there are itemized elements (numbering from 1 to 9 amplifying the components of the standard). For each classroom observation, teachers receive either a "yes" or a "no" on each individual element (28 for Tier I teachers and 29 for Tier II teachers). The Aritfacts and Evidence and Teacher Leader standards were not included in the analysis as they do not lend themselves to video analysis. This information was recorded on the Tier I or Tier II Teacher Observation Forms. The observation forms are retained and used as a main source of data in calculating the final evaluation rating score each teacher receives at the conclusion of the school year.

The observation process in the Clarke County School District Teacher Evaluation System is similar for both Tier I and Tier II teachers. Namely, each teacher will have a pre-observation conference (though the pre-observation conference is only strongly suggested for Tier II teachers), a classroom observation, and a post-observation conference. The major difference between the tiers is the number of times each teacher goes through this process. In Tier I, the teacher will be observed a minimum of three times, two of which may be unannounced. For Tier II teachers, the observation cycle occurs twice per year, with one observation being announced and one being unannounced. For all teachers, additional observation cycles may be performed at the discretion of the evaluator and may be either announced or unannounced.

One of the main needs of the Clarke County School District Teacher Evaluation System included validation of the observation instruments by determining how effectively the system was at identifying teacher quality across the four potential rating categories. This analysis is critical as past research has shown that the ability to accurately identify teacher quality has been difficult to do (Stodolsky, 1984). This need to validate the Clarke County School District Teacher Evaluation System provided the ideal situation to outline the validation process in this specific educational context.

Research Questions

There are four main research questions which guided this study. Questions which this study sought to answer included:

1. What are the types of validity evidence required to illustrate that the observation instrument under examination is a valid representation of teacher quality?

2. Is the inter-rater reliability coefficient acceptable for the observation instrument, and is the paradox which often impacts variations of the Cohen's kappa statistic evident in this teacher evaluation context?
3. Should the paradox be evident, is it more evident with higher-rated or lower-rated teachers?
4. Does Gwet's AC1 statistic (2002) provide a more stable measure of inter-rater reliability than does Fleiss' kappa (1971) in a teacher evaluation context?

Overview of the Research Procedures

The methods used in this study were primarily quantitative in nature, though it is noteworthy that some of the validation evidence gathered outside of the scope of the present study came from qualitative methods and processes. For the present study, the inter-rater reliability coefficients reported, namely the percent of overall agreement, Fleiss' kappa (1971), and Gwet's AC1 statistic (2002), were all calculated using AgreeStat, version 2011.3. The data for this study came from one primary source.

The data for this research came from a validation study conducted in May, 2013 in which four videos of teachers teaching a lesson during the 2012-2013 school year were examined. While ideally this instrument would have been piloted prior to its implementation to determine its validity, this was not completed. As such, this validation study was designed to meet this purpose. Each of these four teachers is an employee of the same school district as were each of the administrators charged with rating the videos, as the calculation of inter-rater reliability requires that "two or more observers have rated the same set of observable evidence" (Graham et al., 2012, p. 13). Those participating in the videos included two elementary school teachers, one middle school teacher, and one

high school teacher. Each of the videos was rated by up to 41 principals and assistant principals in the Clarke County School District, and each of the videos was between 30 and 40 minutes in duration.

Similarly, the principals and assistant principals rating the videos also come from the elementary, middle, and high schools within the district. The ratings were completed with a strict protocol which was explained to each of those participating in the study (see Appendix A). The inter-rater reliability analyses included calculations of all of the participating teachers combined, as well as analyses for each teacher, individually.

Significance of the Study

The significance of this study is three fold. First, illustrating the procedures used to validate and to establish the reliability of a teacher evaluation observation instrument is of use to local educational authorities (LEAs) which have undertaken the process of creating new teacher evaluation systems. Not only does this study illustrate procedures for establishing validity, but this study also suggests the types of validation evidence that are appropriate for situations similar to the one presented in this study.

Second, teacher evaluation data are used to determine if and to what degree the paradox which often impacts kappa statistics exists in an educational context. With the increased emphasis on assessing the quality of teachers through the observation process, establishing the reliability of the instruments used is vital. Should there be problems with the often reported statistics used to establish the reliability of the instruments, the validation of the instruments is difficult to complete. Using these data, the problem of the paradox is further examined to determine if higher-rated or lower-rated teachers are

more impacted by this paradox. Third, a more robust estimate of inter-rater reliability is examined to determine how well it functions in a teacher evaluation context.

Assumptions

The following assumptions were made in completing this study:

1. The principals and assistant principals completing the teacher evaluation instrument for each of the teachers in the videos did so in a consistent manner and with adequate training.
2. Personal relationships between teachers and evaluators were not a factor in the rating of the teachers participating in the study.
3. Principals and assistant principals completed the rating of each of the teachers alone and without assistance from one another.

Definition of Terms

The following terms are defined as they provide background knowledge about the purposes and intents of this study.

Teacher Quality – "Quality teaching could be understood as teaching that produces learning. In other words, there can indeed be a task sense of teaching, but any assertion that such teaching is quality teaching depends on students learning what the teacher is teaching" (Fenstermacher & Richardson , 2005, p. 186).

Validity – "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (AERA, APA, & NCME, 1999, p. 9).

Inter-rater Reliability – "a measurement of the consistence of the *absolute value* of evaluators' ratings" (Graham et al., 2012, p. 5 emphasis in the original)

Fleiss' kappa – a statistical measure designed to assess the reliability of agreement between a fixed number of raters greater than two. This is a variation of Cohen's kappa (1968) which is designed to assess the reliability of agreement between two raters (Fleiss, 1971).

Gwet's AC1 Statistic – a statistical measure designed to assess the reliability of agreement between multiple raters which is a "paradox-resistant alternative to the unstable kappa coefficient" (Gwet, 2012, p. 70).

Study Limitations

The study was limited by the following:

1. The teachers used in the study were purposefully selected as meeting certain criteria and were not selected in a randomized way.
2. As the researcher had only a four hour block of time with the group of principals and then with the assistant principals, the maximum number of videos which could be used in the study was four.
3. While teachers were selected at the elementary, middle, and high school levels, only one teacher was chosen at the middle and high school levels. While this is a limitation, these numbers were selected so as to cover each level of schooling while being somewhat proportional to the overall number of elementary (14), middle (4) and high (2) schools in the district.

Organization of the Dissertation

Chapter 1 outlines the rationale for conducting the study, including a discussion of both the study's significance and the research questions which the study examined. Chapter 2 includes a discussion of the relevant literature related to teacher evaluation

systems, specifically the observation process and the instruments used therein, validity, and the literature related to inter-rater reliability, with an emphasis placed on both kappa, and its variations (Cohen, 1968; Fleiss, 1971), and Gwet's AC1 (2002) statistics, specifically.

Chapter 3 describes the methodological procedures used to complete the study, including the framework of the study, a more in-depth discussion of the instruments used in the Clarke County School District Teacher Evaluation System, the population under study, the data collection procedures, and the analyses procedures used. Chapter 4 discusses the statistical analyses and the findings from these analyses. Finally, Chapter 5 includes the discussion of the major findings, study limitations, implications for policy makers, implications for the field of practice, directions for future research, and final perspectives gained from this study.

CHAPTER 2

REVIEW OF RELATED LITERATURE

Introduction

The purpose of this study was two-fold. The first purpose was to outline validation and reliability procedures in a teacher evaluation context so that observation instruments can be rigorously assessed for quality. The second purpose of this study was to determine if the typically reported measures of inter-rater reliability are adequate when assessing inter-rater reliability in an educational context with multiple raters.

The questions that guided this study were:

1. What are the types of validity evidence required to illustrate that the observation instrument under examination is a valid representation of teacher quality?
2. Is the paradox which often impacts variations of the Cohen's kappa statistic evident in this teacher evaluation context?
3. Should the paradox be evident, is it more evident with higher-rated or lower-rated teachers?
4. Does Gwet's AC1 statistic (2002) provide a more stable measure of inter-rater reliability than does Fleiss' kappa (1971) in a teacher evaluation context?

This study is timely because, in response to the conditions about teacher quality set forth in RTTT, teacher evaluation systems across the country are being revamped. These newly revamped teacher evaluation systems, and the observation instruments used within,

must be validated so that the information gathered from these systems can be trusted.

The primary research methods employed in this study were quantitative; however, some background validation processes and procedures used to develop the instrument came from qualitative sources. This chapter presents the relevant literature across four areas which are of interest to this study: 1) the accountability movement as it relates to teacher evaluation, 2) the purposes and intents of, and problems with previously used teacher evaluation systems, 3) validity and types of validity evidence, and 4) inter-rater reliability and relevant methods for its calculation.

The Accountability Movement and Teacher Evaluation

The beginnings of the field of teacher evaluation as practiced today have its roots in the early 1980s, when the educational reform movement began in the United States with the publication of the educational report, *A Nation at* Risk, in 1983 (Wise, Darling-Hammond, McLaughlin, & Berstein, 1984). As a result of this report, many new laws were created specifically aimed at promoting educational excellence through the identification of both underperforming students and less than able teachers (Tyack & Cuban, 1995).

Of particular importance to the current climate surrounding teacher evaluation is the remedy proposed to deal with the issues outlined in *A Nation at Risk*, which were "more discriminating standards for evaluating and compensating teachers" and "more standardized testing of pupil achievement" (Tyack & Cuban, 1995, p. 79). Additionally, the report argued that "salary, promotion, tenure, and retention decisions should be tied to an effective evaluation system that includes peer review so that superior teachers can be

rewarded, average ones encouraged, and poor ones either improved or terminated" (National Commission on Excellence in Education, 1983, p. 30).

The changes impacting the field of teacher evaluation during this time were three-fold. One, there was an increased focus on accountability as teacher evaluation was increasingly tied to various student outcome measures. Two, knowledge about teacher evaluation and effective practices in assessing teacher quality grew tremendously in the 1980s (Ellett, 1980, 1987; Joint Committee on Standards for Educational Evaluation, 1988; McLaughlin & Pheifer, 1988; Darling-Hammond & Millman, 1990; Scriven, 1988; Wise et al., 1984). Finally, there was increased political pressure to attack shortcomings in the educational system through teacher evaluation as teachers were regarded as the most direct means to effect student achievement because of their direct contact with students (Ellett & Teddlie, 2003).

Throughout the 1990s, reforms to the teacher evaluation process continued. One of the primary changes to the teacher evaluation process during this time, the second major historical shift in teacher evaluation, was a shift away from focusing during the classroom observation on teaching and teaching methods to one focused more on the student and his/her learning (Ellett & Teddlie, 2003). This shift was paramount to laying the foundation for the federal legislation entitled No Child Left Behind (NCLB) in 2001. NCLB is federal legislation geared at improving education through the use of standards-based instruction. This legislation encouraged accountability by setting high standards for both teachers and students and also by creating measurable objectives to determine if the standards were being met (Marsh & Willis, 2007).

The NCLB legislation placed the focus of the accountability movement directly on student achievement in all aspects of educational evaluation, and the evaluation of teachers is no exception. In fact, the Race to the Top (RTTT) federal grants initiative in 2009 has, as one of its main foci, a direct emphasis on being able to identify more accurately quality teaching through teacher evaluation, primarily by linking these data to student achievement data through the use of value-added models (VAMs). For clarity, VAMs have been defined as "methods of analyzing gains, growth in scores, or the amount of knowledge added from year to year as students progress through school" (Amrein-Beardsley, 2008, p. 65).

Related to student achievement, the discussion in this chapter has centered primarily on the history and evolution of the teacher evaluation process. While this is foundational to understanding the current climate of accountability within the educational field as a whole, and teacher evaluation specifically, it is important to also briefly discuss the evolution of the purpose and uses of student achievement data added to the teacher evaluation process to understand better how teacher evaluation is a politically charged topic in the current educational environment in which the Federal Government has by fiat put controls on teacher and leader evaluation systems, processes, and procedures as a way to increase accountability.

Linking student achievement data to data from the evaluation of teachers has been commonplace since the 1970s and 1980s (Bolton, 1972; House, 1973; Jaeger & Tittle, 1980; Kleinman, 1966; Medley & Coker, 1987; Popham, 1971); however, using this relationship for purposes of personnel decisions is a much newer development. These early studies addressed the relationship between student achievement data and teacher

evaluation; however, these studies were more research-focused and not intended for use in personnel decisions. In fact, these studies often failed to discover a significant link between teacher evaluation and student achievement, making the use of such data for personnel decisions "bad practice" (Koedel & Betts, 2011; McCaffrey et al., 2009; Newton et al., 2010; Papay, 2011; Sass, 2008).

As discussed, the accountability movement, which has roots in the 1980s, has grown into the most recent piece of government intervention into public education with the implementation of the 2009 federal grants program, Race to the Top (RTTT). In this initiative, states and school districts across the country are required to include measures of student achievement into the evaluation of teachers (Hanover Research, 2011). It is the inclusion of student achievement data into the evaluation process in what may be a punitive way that makes it different from prior uses of such data. The use of these data in such a way is unsupported in the literature (Koedel & Betts, 2011; McCaffrey et al., 2009; Newton et al., 2010; Papay, 2011; Sass, 2008). While these data on student learning should be considered in the teacher evaluation process, it should be only one factor that should be considered among the many comprising what it means to be a teacher and should be avoided in high-stakes personnel decisions (McCaffrey, Lockwood, Koretz, & Hamilton, 2003).

To address the requirement of RTTT, several states have begun the process to include estimates of the value a teacher adds to his/her students into the evaluation process, which has been typically accomplished through the use of value-added models (VAMs). As previously discussed, these models focus on changes in standardized test scores, often failing to include the various other indicators of student growth. Overall,

methodological issues related to the use of VAMs have been clustered into five areas, which are 1) issues of validity (Koedel & Betts, 2011; McCaffrey et al., 2009; Newton et al., 2010; Papay, 2011; Sass, 2008), 2) issues of missing and too little data (Koretz, 2008; McCaffrey & Lockwood, 2011), 3) questions of which demographic and school level variables to include into the analysis (McCaffrey et al., 2003; Tekwe et al., 2004), 4) issues about the stability of measures across time and assessments (McCaffrey et al., 2009; Newton et al., 2010; Sass, 2008), and 5) issues of randomization (Corcoran, 2010; Kupermintz, 2003; Linn, 2008; Rubin, Stuart, & Zanutto, 2004).

While there is no defined number of studies which would suggest that VAMs are valid methods to make claims about teacher effectiveness, the types and severity of decisions made from these methods dictate the type and amount of validation evidence required (Kane, 2009, 2013; Messick, 1989, 1995; Zumbo, 2009). As Kane (2013) argued, "the kinds of evidence required for validation are determined by the claims being made, and more ambitious claims require more evidence than less ambitious claims" (p. 2). Using teacher evaluations to make decisions about retention, promotion, or possibly dismissal are certainly ambitious but serious enough to require stringent validation evidence. What is clear when discussing existing validity evidence related to the use of VAMs is that many studies have shown them to be problematic, indicating a lack of support for continuing their use in high-stakes situations (Koedel & Betts, 2011; McCaffrey et al., 2009; Newton et al., 2010; Papay, 2011; Sass, 2008).

An accurate estimate of the value a teacher adds to his/her students requires longitudinal data which are mostly complete and contain adequate measures to determine student growth. While this may be the goal for VAM analysis, it is often the case, as is

with a great deal of educational data, that there are significant amounts of missing data (Koretz, 2008; McCaffrey & Lockwood, 2011). It is also important to note that the missing data are often different from the data that are complete (Koretz, 2008). Should instances of missing data center on certain teachers, this could be a problematic issue for analysis. Additionally, teachers with too few students (i.e., special education teachers or English language specialists) can be difficult to assess with many of the currently tested VAMs. As Koretz (2008) explained, "all teachers will sometimes appear more or less effective than they really are because of sampling error, and substantially incorrect estimates will be much more common among teachers with smaller classes" (p. 26). While each of these issues of too little or missing data are typically addressed using data simulation techniques, justifying this practice when someone's continued employment may be in doubt, is a very difficult task to undertake.

Thirdly, and of great importance, is whether or not certain variables should be included into the VAM to control for student background factors outside of the teacher's control (McCaffrey et al., 2003; Tekwe et al., 2004). These variables often relate to race and poverty, both individually and at the school level, and have been dealt with in various ways. One system, which is the most widely used VAM model, the Education Value-Added Assessment System (EVAAS), does not control for variables such as these while other programs piloted in various states use methods to control for these types of variables (Hershberg, 2005; Newton et al., 2010). Whether these variables are included in the VAM or not, further research addressing this area is needed. As McCaffrey et al (2003) underscored, "the importance of modeling student background characteristics

when using VAM to estimate teacher effects remains an empirical question that must be addressed" (p. 70).

The fourth area about the use of VAM that must be examined is the stability of the estimate across years and across student achievement measures. Certainly it can be argued that if the VAM estimate is useful and valid in making decisions in such high-stakes areas as teacher retention, promotion, or dismissal, then these estimates must be accurate regardless of the test used. Additionally, and for similar reasons, these estimates must be stable across multiple years of data as well. Studies examining the stability of VAM estimates across time have reported mixed results. McCaffrey et al. (2009) found moderate correlations across different years of value-added teacher rankings in elementary and middle school math teachers while VAM estimates were found to be highly unstable in a study conducted by Sass (2008). Newton et al. (2010) also found that teacher ratings varied across courses and years.

The issue of stability across years is yet unresolved in the literature and needs further examination (McCaffrey et al., 2009; Newton et al., 2010), though the process is based on vertical scaling, which places tests across years on the same scale, and many examinations were not designed to be vertically scaled. Additionally, empirical evidence about the stability of VAM estimates across measures of student achievement is minimal, though Papay (2011) determined that VAMs fail to consistently estimate teacher effectiveness based on which measure of reading achievement was used.

To fairly assess the value a teacher adds to the learning of a student is a very difficult task, in part, due to the non-random way in which students are assigned to teachers, students are assigned to schools, and teachers are assigned to and within the

schools in which they teach. Due to this lack of randomization, the other factors (e.g., issues of missing data, stability of estimates) are further complicated because they vary across teachers and schools, creating additional challenges to making inferences about teacher quality based on student achievement data (Corcoran, 2010; Kupermintz, 2003; Linn, 2008; Raudenbush & Willms, 1995; Rubin et al., 2004).

Needless to say, there are many VAMs currently being developed, and each has methodological concerns, which suggests two things. First, these models are still being developed and improved and should be used sparingly and cautiously in terms of making personnel decisions (Newton et al., 2010; Winters & Cowen, 2013). As Newton et al (2010) argued:

> Despite its conceptual and methodological appeal, the use of VAM to estimate teacher effectiveness or to rank teachers for high stakes purposes poses daunting challenges stemming from many factors: the non-random assignment of students to teachers and schools, the limitations of particular tests both for measuring the full range of desired knowledge and skills and for measuring learning gains, and the difficulties of disentangling the contributions of many influences on learning – multiple teachers, parents, tutors, specific curricula, and the availability of useful learning materials, as well as other school resources like time, class sizes, and the like. (p. 4)

Second, a method is needed to bridge the gap until these methods are further improved, or until better methods are created to include student achievement data into the evaluation of teacher effectiveness.

Not only does the research community voice opposition to the use of these estimates in high-stakes situations, but also professional teacher associations also caution against the use of a single measure as a means to assess overall teacher quality (National Council of Teachers of English, 2012; National Council of Teachers of Mathematics,

2011). Using a sole measure for assessing teacher quality is problematic, especially when decisions of continued employment are based on such estimates.

The profession of teaching involves multiple responsibilities and the abilities to adjust to varying needs of individual students and classes, and thus, should require multiple sources of information to assess the quality of teaching. Not only is assessing this through the use of VAMs nearly impossible because of the lack of attention focused on other aspects of teaching, but also assessing the various qualities of teachers is also problematic because of the limited number of observations typically included in teacher evaluation systems (Zepeda, 2012). Valid conclusions about teacher evaluation require an assessment of factors in multiple rating categories, time to adequately observe and assess teachers, and the use of multiple sources of evidence of quality teaching (Recesso & Zepeda, 2006; Zepeda, 2012).

The previous discussion of how teacher evaluation has changed through the last several decades is meant to illustrate why teacher evaluation systems across the country have undergone changes and will continue to do so as legislated by RTTT. These changes should be undertaken with the purposes and intents of teacher evaluation in mind.

Teacher Evaluation: Purposes, Intents, and Problems

There are two main purposes of teacher evaluation. The act of evaluating a teacher should serve a summative function (evaluation) and a formative function (professional development) (Danielson & McGreal, 2000; Marzano, 2012). The summative function refers to the purpose geared at identifying quality teaching for personnel decisions (McColskey & Egelson, 1993; Mead et al, 2012) and the formative

function refers to the purpose geared at identifying professional needs of teachers so as to improve the quality of teachers individually (Danielson & McGreal, 2000; Mead et al., 2012; Stronge & Tucker, 2000; Zepeda, 2012).

The two main purposes of the teacher evaluation process are generally agreed on in the literature (Danielson & McGreal, 2000; Marzano, 2012). While this is the case, problems with each of the purposes have been identified (Danielson & McGreal, 2000; Loup et al., 1996; McColskey & Egelson, 1993; McGreal, 1983; Wise et al., 1984; Zepeda, 2006, 2013).

The problems with current teacher evaluation systems center around three main issues. First, recently used teacher evaluation systems are often outdated, fail to address all aspects of what it means to be a quality teacher, and were designed around learning concepts that are no longer considered mainstream in the literature (Danielson & McGreal, 2000, 2005; Loup et al., 1996; McGreal, 1983; Wise et al., 1984). In fact, Danielson and McGreal contend that "many evaluation systems in use today were developed in the early to mid-1970s and reflect what educators believed about teaching at the time" (2000, p. 3). Second is that data gathered from these evaluation systems are without much value as they are both from systems which are not valid and are from observations completed by evaluators with little training (Kauchak, Peterson, & Driscoll, 1985; Medley & Coker, 1987; Peterson, 2000; Stodolsky, 1984; Wise et al., 1984). Finally, the data are often used (or not used) in ways which fail to make a meaningful impact on teaching and learning (Weisberg, Sexton, Mulhern, & Keeling, 2009).

Teacher evaluation systems used today are often outdated. Using these outdated systems has led to much previous research which has shown that the evaluation of

teachers by principals has been inadequate both in differentiating between more and less proficient teachers and in guiding teachers to improve teaching ability (Danielson, 1996; Medley & Coker, 1987; Peterson, 2000).

The data gathered from teacher evaluations are problematic and have offered little to improve the quality of teaching and learning (Peterson, 2000; Weisberg et al., 2009). The problems which are most impactful to teacher evaluation data, in examining the literature, center around three issues. First, teacher observations, which are a large part of any evaluation system, are completed too infrequently and for too short of a duration (Zepeda, 2012). Second, many teacher evaluation instruments cannot pass a validity challenge because they fail to completely assess all areas of what it means to be a teacher (Haefele, 1993). Finally, research has shown that the data gathered from teacher observations, completed by principals, were found to be unreliable (Stodolsky, 1984) primarily because principals lack the knowledge required in regard to teacher evaluation, the observation process, and the accurate rating of teachers (Kauchak et al., 1985; Medley & Coker, 1987; Peterson, 2000; Stodolsky, 1984; Wise et al., 1984).

Many of the issues associated with the lack of reliability of data obtained from teacher evaluation systems lie in the systems themselves. What is clear is that, according to Gallagher (2004), "to determine the effects of high-quality teaching, a valid and reliable method of identifying quality instruction is necessary" (p. 80). This has been the crux of the problem in using data from teacher evaluation systems of the past. These systems fail to both pass the validity test (Haefele, 1993) and to provide data which are useful in improving teaching and student learning.

A particularly troubling study examining both the ratings of teachers and uses of teacher evaluation data was published in 2009 by Weisberg et al. In this study, the authors examined school districts, both large and small, and found that nearly all teacher evaluation scores were found to be good or great, excellence in teaching was not rewarded by districts, professional development was rarely tied to the results of the evaluation, new teachers were generally rated above being satisfactory, and negative results of a teacher evaluation rarely led to dismissal. These findings are striking examples of the many problems associated with teacher evaluation as practiced today, and provide further examples of why research and professional associations fail to support using such systems for personnel decisions (Danielson & McGreal, 2000; National Council of Teachers of English, 2012; National Council of Teachers of Mathematics, 2011).

To address these problems, both of using outdated systems and gathering and using data in ways to improve teaching and student learning, it is clear that well-designed teacher evaluation systems, containing new and more exhaustive instruments accurately measuring teacher quality are necessary. These instruments should, at a bare minimum, be used frequently enough and for long enough in duration to get an accurate assessment of teacher quality, accurately measure all aspects of quality teaching, and should be completed by evaluators with the required training and knowledge to differentiate between levels of teaching (Haefele, 1993). Additionally, these instruments should be created with input from all stakeholders (i.e., principals, district personnel, teachers, etc.), and in an ideal world, serve as a piece of the overall evaluation and not as its entirety (Goe & Holdhelde, 2011). Multiple measures of teacher effectiveness, as argued in the

literature, provide the most accurate assessment of teacher quality (Goe & Holdhelde, 2011; Recesso & Zepeda, 2006; Zepeda, 2012).

To create more meaningful teacher evaluation systems, thereby creating impactful evaluation data, teacher evaluation systems must, through the evaluation instrument, more accurately identify the various levels of teaching. There are several steps to bring about this outcome. First, instruments must be created so as to have more than the typical dichotomous ratings used in many of the existing systems (Danielson & McGreal, 2000). The ability to rate teachers across multiple rating categories (i.e., Exemplary, Proficient, Emerging, or Unsatisfactory) helps to assist possibly in more accurate identification of teacher quality (Danielson & McGreal, 2000).

To address the criticism that the validity of the evaluation system is not assured because the instrument fails to address all aspects of quality teaching suffered by many current evaluation system instruments, a new evaluation system needs to be created so that teachers are rated on *each* item deemed important to quality teaching (Danielson & McGreal, 2000). It is impossible to measure accurately teacher effectiveness if all of what defines teacher effectiveness is not addressed. Additionally, what defines teacher effectiveness should be established in concert with the stakeholders involved, including teachers (Goe, Holdhelde, & Miller, 2011).

Addressing the reliability and validity issues, both through multiple ratings and by creating a system that accurately covers aspects of teaching, will create valuable data for both assessing teacher effectiveness and as a source to improve teaching through professional development. Weiner and Jacobs (2011) summarized, "creating information

that is credible and useful for developing teacher effectiveness should be treated as important priorities alongside technical concerns like validity and reliability" (p. 5).

Validity and Validity Evidence

One of the main questions guiding this research is "what are the types of validity evidence required to illustrate that the observation instrument under examination is a valid representation of teacher quality?" and is therefore relevant to this section of the review of the literature. In the *Standards for Educational and Psychological Testing* (1999), "validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (AERA, APA, & NCME, 1999, p. 9). It is this definition, along with the work of Kane (1992, 2006) on the argument-based approach to validation, which guided this discussion.

The concept of validity has gradually changed from its inception. Certain models like the content-based model, the criterion model, and the construct model were all created as a means to validate specific types of test uses (Kane, 2013). While each has strengths and weaknesses, the result had fragmented the concept of validity. In response, Messick (1989) argued that these various validity models could all be discussed in terms of a unified model of validity based on content. In fact, Messick (1989), under this unified model, defined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (p. 13, emphasis in the original). This unified model of validity was a seminal moment in the study of validity (Kane, 2013) and much of the work undertaken

since has been completed with this model as a "philosophical underpinning" of the field (p. 49).

Kane's (1992, 2006) argument-based approach to validity, which is based on the Interpretive/Use Argument (IUA), calls for a "clear statement of the proposed interpretations and uses and a critical evaluation of these interpretations and uses" (Kane, 2006, p. 17). This approach to validation has as its foundation a clear and well-thought out statement of the proposed interpretations and uses of a test or instrument. These uses are then evaluated through an assessment of their plausibility. This evaluation is referred to, by Kane (2013), as the validity argument. The extent to which these interpretations and uses are plausible determines if the argument is valid (Kane, 2013).

Various types of validation evidence exist because various instruments (tests, observation instruments, etc.) are found to be either valid or invalid in different ways. Kane argued that "by specifying the claims being made, it provides guidance on the kinds of evidence needed for validation (Kane, 2013, p. 6). Kane (2013) stated:

> Different kinds of warrants require different kinds of backing. Scoring rules that take us from observed performances to a score generally rely on expert judgment about the criteria to be used in scoring and on quality control of scoring procedures (and possibly on data regarding rater accuracy and consistency). Generalizations from a sample of observations to expected performance over domains of possible observations rely on evidence that the sampling was consistent with the statistical model being employed, and on generalizability (or reliability) analyses (or IRT based analyses) indicating that the sample was large enough to control sampling errors. Extrapolations to different kinds of performance in different contexts rely on empirical evidence (e.g., from a regression analysis) and/or on analyses of the overlap in the skills required for the different kinds of performances in different contexts. Theory-based inferences rely on evidence for the theory and for the appropriateness of the test scores as indicators of constructs in the theory. Score-based decision procedures require evidence that the procedure achieves its goals without unacceptable negative consequences. (pp. 10-11)

Of these various types of validation evidence, those most relevant to the current study are those related to scoring rules and score-based decisions.

As is clear in the intended purposes of the Clarke County School District Teacher Evaluation System, the system is designed to assess accurately teacher quality, to be able to better discriminate among teachers within the district through a scoring system based on 4 score points, and to use these data for retention/promotion decisions, and to support teacher professional development. With this in mind, and after assessing the content through expert judgment, the additional evidence needed to either validate or to invalidate the observation instrument center around the quality control of the scoring procedures, rater accuracy and inter-rater reliability, and an assessment of the potential unacceptable negative consequences relevant to the instrument.

Evidence of the quality control of the scoring procedures used in the study is outlined in Chapter 3. For this study, the scoring procedures were well defined (see Chapter 3) and the conditions set forth for the study were uniform throughout the process, with the exception of time of day in that the validation study occurred in the mornings for the principals and in the afternoon for the assistant principals. Whereas, classroom observations occurred at the site levels all throughout the day in classrooms where teachers and students were "live" and in the validation process, digital recordings were used to record classroom observations.

In terms of potential unacceptable negative consequences, the types of potential validity evidence are primarily qualitative in nature and rely heavily on expert opinion. Sources of evidence include an expert assessment of the potential negative consequences

and a determination if any of these consequences are deemed significant enough to justify the elimination of the instrument as constructed.

Inter-rater Reliability and Methods of Calculation

The following discussion is related to both inter-rater reliability and methods of calculation. The remaining three research questions framing this study drive the following discussion of relevant literature. In situations such as the one presented in this study, one of the main sources of validity evidence is inter-rater reliability, as without establishing a consistency of ratings, valid uses of the scores from the observation instrument cannot be made . Inter-rater reliability, in this study, refers to "a measurement of the consistence of the *absolute value* of evaluators' ratings" (Graham et al., 2012, p. 5, emphasis in the original).

Inter-rater reliability is a technique for determining the consistency of raters when tasked with accurately assessing what they have seen. There are a number of statistical means of evaluating inter-rater reliability, and some of the more common methods used include percentage of exact agreement, Cohen's kappa (1968) and its variations, and the intra-class correlation coefficient (Cook & Beckman, 2006; Graham et al., 2012). While these methods are common, additional methods, such as Gwet's AC1 statistic (2002) have also been created as a means to assess inter-rater reliability.

As Graham (2012) et al. argued, "the percentage of absolute agreement is the simplest to understand" (p. 7). The percentage of absolute (referred to as exact in this study) agreement is simply a count of the number of times raters agree divided by the total number of ratings. While this measure is straightforward, the percentage of exact agreement does not account for chance agreement.

Cohen (1968) created a kappa statistic which "corrects for the likelihood that some agreement between evaluators will occur by chance" (Graham et al., 2012, p. 7). Cohen's kappa (1968) was designed to be used with two raters, though variations such as Fleiss' kappa (1971) have been created to be used in situations with multiple raters. It is Fleiss' kappa (1971) which is reported in this study.

The final common method reported in inter-rater reliability studies is the intra-class correlation, though this statistic is of no interest to this study as it "is a measure of agreement that is useful when there are many rating categories (5 or more) or when ratings are made along a continuous scale" (Graham et al., 2012, p. 7). Neither of these conditions, which make the intra-class correlation of interest, applies to this study.

Reporting the percentage of exact agreement and Fleiss' kappa (1971) are clearly outlined in the literature, problems with variations of kappa have been noted in other research (Feinstein & Cicchetti, 1990; Graham et al., 2012; Gwet, 2002, 2012). It has been noted that kappa statistics suffer from a paradox where high values of overall agreement can produce low values of kappa (Feinstein & Cicchetti, 1990; Gwet, 2002, 2012). In fact, Gwet (2012) argued that "kappa often yields coefficients that are unexpectedly low if compared to the overall percent agreement" (p. 36). In response to this paradox, Gwet (2002) created Gwet's AC1 statistic which is designed to function similarly to and to be interpreted as a variations of kappa, but without suffering from the same paradox.

A brief discussion of the methods of calculation for both Fleiss' kappa (1971) and Gwet's AC1 statistic (2002) is warranted. The formula for Fleiss' kappa (1971) is

$$\frac{\bar{P} \quad \bar{P}}{\bar{P}},$$

where the numerator provides the observed agreement above chance and the denominator provides the maximum possible degree of agreement above chance, as $\overline{P}$e is the estimated chance agreement and $\overline{P}$ is the observed agreement .

Similarly, Gwet's AC1 statistic (2002) is calculated as

———,

where the numerator provides the observed agreement above chance and the denominator provides the maximum possible degree of agreement above chance, as e(y) is the estimated chance agreement and p is the observed agreement.

The main difference between Gwet's AC1 statistic (2002) and variations in kappa statistics lies in the chance corrected part of the equations, namely the $\overline{P}$e part of the equation for Fleiss' kappa (1971) and the e(y) part of the equation for Gwet's AC1 statistic. The chance corrected part of the equation for Fleiss' kappa (1971) is designed so that is can range from 0 to 1. Gwet's AC1 statistic (2002), on the other hand, is bounded to range in value from 0 to 0.5, which has as foundational the idea that the maximum chance agreement between two raters, when one randomly guesses, would be 0.5. Bounding the upper value of the estimated chance agreement to 0.5 is what results in Gwet's AC1 statistic (2002) being paradox resistant. Fleiss' kappa (1971), having an upper bound of 1 for the estimate of chance agreement results in the possibility of an inter-rater reliability value of zero.

Both Fleiss' kappa (1971) and Gwet's AC1 (2002) are statistics designed to account for the fact that agreement between raters can occur by chance, though they differ in the way that the chance of agreement is calculated (Gwet, 2012). According to Gwet (2012), the kappa coefficient relies on an improbable assumption that "all ratings

are known to be independent even before the experiment has been carried out" (p. 70), though raters regularly rate the same subjects, often producing ratings that are dependent. Gwet's AC1 (2002) statistic, on the other hand, relies on the "more realistic assumption that only a portion of the observed ratings will potentially lead to agreement by chance" (Gwet, 2012, p. 70), which makes this a more reliable, paradox-resistant inter-rater reliability coefficient (Gwet, 2002, 2012; Wongpakaran, Wongpakaran, Wedding, & Gwet, 2013).

Each of the inter-rater reliability statistics presented in this study ranges in value from 0 to 1 (Graham et al., 2012; Gwet, 2002). The interpretation of the statistics, however, can be difficult. In fact, the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999) do not suggest a criterion for the interpretation of these measures, only that they should be calculated and reported.

There do exist, however, a criterion to assist with the interpretation of these statistics reported in the literature. Graham et al. (2012) have made suggestions for the sufficiency of the statistics and the consequential use of the ratings. For the percentage of exact agreement, a high value is 0.9 and a minimum value is 0.75. For Cohen's kappa (1968), and its variations, a high value is 0.81 and a minimum value is 0.61. Graham et al. (2012) make no recommendations for values of Gwet's AC1 (2002) statistic, though the interpretation is similar to that of Fleiss' kappa (1971) (Gwet, 2002) and thus the same guidelines outlined above were used.

The relevant literature discussed above is meant to lay the groundwork for the chapters to come. The methods employed in conducting this study are discussed in Chapter 3, including a discussion of the procedural steps in creating the observation

instrument under examination and the statistical methods employed to assess inter-rater reliability as part of the validation of the instrument.

CHAPTER 3

RESEARCH METHODS

This chapter discusses the various methods used in this study to gather the evidence needed to validate the rating instrument as it was constructed in the Teacher Evaluation System in the Clarke County School District. In addition, this chapter outlines the context in which this study occurred, the raters participating in the study, and the rating instrument.

The methods employed in this study were primarily quantitative in nature, though some of the evidence gathered for validation purposes came from qualitative sources to examine the first research question of the study. As previously discussed, the evidence gathered to determine if the observation instrument is valid in its intended use is centered on two primary sources. One, evidence suggestive of the degree to which the construct of teacher quality was exhaustively covered, as well as procedural information on the construction of the instrument was gathered. The second main source of validation evidence in situations in which rating instruments are under examination is evidence of inter-rater reliability.

Chapter 3 begins by setting the context under which the study takes place by providing an overview of the Clarke County School District. Following this description of the context, information pertaining to the observation instrument and the raters participating in the study is presented. Third, the chapter then outlines the data and the method by which these data were collected. Finally, the chapter concludes with a

discussion of the methods used to calculate measures of inter-rater reliability and potential problems with typically used measures of inter-rater reliability.

Overview of the Clarke County School District

Located in Athens, Georgia, the Clarke County School District serves 12,750 students. Ethnically, 51% of the students are African-American, 23% are Hispanic, 20% are white and 2% are Asian. Nearly 12% of the students have English as their second language and 11% are special needs students. There are 2,341 employees—58% white, 38% African-American and 3% Hispanic. Students benefit from the expertise of nearly 1,200 teachers—with over 60% having advanced degrees and over 250 certified in gifted education. There are 16 National Board Certified teachers and 7 Georgia Master Teachers.

As a community, Athens-Clarke County, has the third-highest poverty rate among U.S. counties with populations between 65,000 and 249,000. Over 30% of children in Clarke County live in poverty—around 82% of students are eligible for the federal meal program, 49% live in single-parent homes, and 19% of adults (>25 years) did not complete high school—all significant risk factors that can keep students from graduating from high school. Despite these factors, the district's current graduation rate is 70.8, which is above the state average. The Clarke County School District was named the 2010 Title I Distinguished School District (Large) for reducing the achievement gap between economically disadvantaged and non-disadvantaged students.

The Observation Instrument

In 2012, the Clarke County School District Teacher Evaluation System was implemented. As part of this system, the Tier I and Tier II Teacher Observation

Instruments were created. Teachers are placed into one of two tiers, based on experience and performance. Tier I teachers include all teacher in years 1, 2, or 3, teachers new to the school district, and teachers placed on a formal plan of improvement. Tier II teachers include all teachers in years 4, 5, or 6 and experienced teachers new to the district (at the principal's discretion). For this study, Tier I and Tier II classroom observation instruments were evaluated, and these analyses are exactly the same as the standards which can be evaluated through video evidence are consistent across each of the instruments.

Teachers in the Clarke County School District are rated, based on their classroom observations, on either 6 (Tier I) or 7 (Tier II) performance standards which have been agreed on by the system as covering the requirements of being a high quality, professional educator in a standards-based classroom. The standards for Tier I included *Curriculum and Planning (CP), Standards-Based Instruction (SBI), Assessment of Student Learning (ASL), Instruction Environment (IE), Building Positive Student Relationships (BPSR) and Artifacts and Evidence (AE)*. Tier II teachers are assessed on the same 6 standards as Tier I teachers, though there is an additional standard, *Teacher Leader (TL)*, which asserts that as a teacher becomes a more seasoned veteran educator, they have a responsibility to assume additional responsibilities benefitting the school community and/or their peers. For this study, both the *Artifacts and Evidence* and the *Teacher Leader* standards were not assessed.

Within each of the performance standards, there are itemized elements (numbering from 1 to 9 amplifying the components of the standard). For each classroom observation, teachers receive either a "yes" or a "no" on each individual element (28 for

Tier I teachers and 29 for Tier II teachers). For purposes of this study, only 25 elements were assessed across all of the participating teachers.

Raters

The raters participating in the study included 42 principals and assistant principals in the school district. While others within the district are certified to rate teachers, only those regularly tasked to do so were included in the study.

Of the 42 participants, 38% are male and 62% are female. The racial composition of the participants is 55% African-American and 45% Caucasian. There are 12 participants with doctoral degrees. Finally, each of the participants received extensive training on the instrument through a two-day training. Additionally, ongoing training was also provided through monthly principal meetings. Finally, each of the raters was provided with copies of the rating rubrics which were created to accompany the instrument.

Data

The data used in this study came from two main sources. The first source of the validation evidence came from primarily qualitative sources. These included existing interview data and school district documents which outlined the procedures followed when the Clarke County School District Teacher Evaluation System was created. Moreover, data related to the content of the teacher observation instruments were examined relative to how the performance standards and the indicators were identified and then verified as high-yield instructional strategies. Artifacts, including observable practices aligned to the performance standards and the indicators of teacher quality as found in the literature and reported by Zepeda were vetted by school leaders, teachers,

and others who conduct classroom observations. Additionally, the results of field-testing were presented to relevant stakeholders and the numerous iterations of the instruments were examined for potential improvement to the instrument.

The second data source was video-taped lessons. These were used for the assessment of inter-rater reliability. The tapes were collected as part of a study designed for the express purpose of providing the school district with the coefficients. The data gathering segment of the study was conducted at the Clarke County School District main office complex and was completed over the course of one day.

The data come from videos of teachers that taught actual lessons within the district during the 2012-2013 school year. There were 15 videos of teachers teaching, four (limited by time constraints) of which were chosen to be used in the study. These videos were chosen in a purposeful way so as to cover all levels of schooling, cover a range of academic subjects, and each be of an appropriate duration.

Of the four teacher videos chosen, two teachers were from elementary schools, one was a middle school teacher, and one was a high school teacher. Elementary school teachers had two representatives because of the 20 traditional schools within the district, 14 were elementary schools. Each of the videos was between 30 and 40 minutes in length. The subjects covered were varied and included science, reading, and mathematics.

The principals and assistant principals across the district all participated in this segment of the study. The principals' session lasted from 8 a.m. until 12 noon and the assistant principals' session lasted from 12:30 p.m. until 4:30 p.m. During the principals' session, a strict protocol was used to conduct the classroom observation simulations (see

Appendix A). After the protocol was read, principals were then each given a copy of the observation instrument (see Appendix B), which included two additional questions, one multiple choice and one short answer. The two additional questions were: 1) If you were using ONLY this observation as a way to assess this teacher, which rating would you give this teacher on his/her overall performance: a. Exemplary, b. Proficient, c. Emerging, or d. Unsatisfactory, and 2) List two or three items you would want to probe this teacher on during the post-observation conference.

Once the principals were given the observation instrument, the first video was shown. At the completion of the video, principals were given 10 minutes to complete the observation form and then were given a 10 minute break. This same procedure was repeated for each of the remaining three videos.

The assistant principals were then asked to go through the same process later the same day. The data were transferred from the paper hardcopies to a spreadsheet file in Microsoft Excel for input to the AgreeStat 2011.3 Program. The dataset was created and 20 percent of the 4050 cells were then checked at random to ensure accuracy. None of the cells that were checked were found to be incorrect.

Related to inter-rater reliability, Cohen's kappa (1968), and the derivations of it, have been shown, through a discussion of related literature, to suffer from a paradox in which high levels of exact agreement can result in artificially low values of kappa (Gwet, 2012). As such, Gwet's AC1 statistic (2002) was used, along with the percent of exact agreement and Fleiss' kappa (1971), as the measures of inter-rater reliability for the classroom observation instrument.

These statistics were chosen for two reasons. One, inter-rater reliability evidence was needed to help with the overall validation of the observation instrument. Two, these exact three measures were needed to complete the second main purpose of this study, which was to determine if the paradox associated with measures of kappa was evident in a teacher evaluation context and to determine if Gwet's AC1 statistic (2002) functioned as a better measure of inter-rater reliability than did Fleiss' kappa (1971).

CHAPTER 4

FINDINGS

This study sought to answer 4 primary questions. These questions were:

1. What validity evidence exists to illustrate that the observation instrument under examination is a valid representation of teacher quality?
2. Is the inter-rater reliability coefficient acceptable for the observation instrument, and is the paradox which often impacts variations of the Cohen's kappa statistic evident in this teacher evaluation context?
3. Should the paradox be evident, is it more evident with higher-rated or lower-rated teachers?
4. Does Gwet's AC1 statistic (2002) provide a more stable measure of inter-rater reliability than does Fleiss' kappa (1971) in a teacher evaluation context?

The results presented here are ordered by research question.

The first research question deals with the content of the observation instrument and the processes involved in enacting classroom observations. The data examined to explore this question came from historic documents; interviews with the expert in instructional supervision and teacher evaluation who assisted the Clarke County School District develop the Teacher Evaluation System, and various artifacts.

First, the procedural aspects of building the classroom observation instrument are examined, and then second, the content of the Classroom Observation Instruments for Tier I and Tier II teachers are examined.

Procedural Validation—Process

Does the evidence to determine whether or not the procedures used to develop the classroom observation instruments used for Tier I and Tier II teachers being studied is a valid representation of teacher quality support the idea that the steps typically followed (e.g., consulting experts) to ensure content validity exist?

The processes undertaken by the Clarke County School District to create the teacher evaluation observation instrument were documented spanning a three year timeframe. System documents were examined and illustrated multiple stakeholder engagement. The major procedural steps are detailed in Appendix C. In short, teachers, leaders, and an advisory board were actively engaged in not only giving feedback but also in pilot testing earlier iterations of the observation instrument, and the evaluation system with which it is associated. Additionally, an outside expert further developed, refined, and pilot-tested various aspects of the instrument throughout years 1 and 2 of this initiative.

Content Validation

Evidence related to the content and the exhaustiveness of the coverage of the construct can be in multiple forms, though the primary source of data of this type is a reliance on expert judgment (Kane, 2013). Expert judgment, as related to the instrument in this study, relied on the knowledge of a teacher evaluation and classroom observation expert, principal input, input from district-level leaders , and input from teachers. An additional source of evidence is a comparison of the standards covered in the observation instrument to those that are accepted as covering the construct in the existing literature.

As is clear, the creation of the observation instrument was done so in a multistep way, was influenced in many steps by expert judgment (including principals, district

personnel, teacher, and a research professional), and relied heavily on peer-reviewed research so as to accurately measure teacher quality. As process, the teacher evaluation and classroom observation expert culled the research and best practices from the literature and adapted and applied these constructs to the classroom observation instruments for the Clarke County School District.

Table 4.1 identifies a sampling of the major studies which influenced the development of the standards contained on the instrument following the performance standards and their elements.

Table 4.1: Sample of Major Studies Influencing the Development of the Standards on the Observation Instrument

Standard 3 Assessment of Student Learning Standard (AL): The teacher uses a balanced variety of assessment techniques that are systematically implemented, resulting in appropriate interventions that foster continuous improvement for all. NOTE: One or more assessment strategies should be observed.

Element	**Look For, Teacher**	**Look For, Students**	**Research Base**
The teacher intentionally solicits feedback from all students on their understanding of the standard.	Feedback from students allows the teacher to check for student understanding of the standard. Does the teacher • Ask students to provide feedback. • Engage students in review of their own work and others. • Monitor and adjust strategies in response to learner feedback. • Vary role in the instructional process (e. g., instructor, facilitator, coach, audience) in relation to the content and purposes of instruction and the needs of students. • Use a variety of clear, accurate presentations and representations of concepts, using alternative explanations to assist students with the standard.	(Jang & Stecklein, 2011) (1) repetition; (2) incorporation; (3) self-repair; (4) peer repair; (5) acknowledgement; (6) same error; (7) different error; (8) off target; (9) hesitation; (10) partial repair (Lyster, 1998a; Lyster & Ranta, 1997; Sheen, 2004) Effective learning results from students providing their own feedback, monitoring their work against established criteria (Trammel, Schloss, & Alper, 1994; Wiggins, 1993) Students can monitor and provide feedback to other students and compare their work to criteria.	(Black & Wiliam, 1998b; Orlander & Fincke, 1994) (Orlander & Fincke, 1994) (Zacharias, 2007) (Jang & Stecklein, 2011) (Lyster, 1998a; Lyster & Ranta, 1997; Sheen, 2004)

These sources assisted with the overall validation of the observation instrument in terms of how well the instrument covers the construct of teacher quality as reported in numerous synthesis of research (e.g., Barge, 2012; Marzano, 2003; Waters, Marzano, & McNulty, 2003).

Inter-rater Reliability

The second question guiding this study was to determine the adequacy of the observation instrument in terms of overall inter-rater reliability and also whether or not the typically reported measures of inter-rater reliability (percent overall agreement and some variation of kappa) are adequate to assess inter-rater reliability in an educational context, or is the paradox which sometimes impacts kappa statistics evident (Research Question 2). The variation of Cohen's kappa (1968) used in this study is Fleiss' kappa (1971), which was created to be used in situations such as this, one with multiple raters. Both Fleiss' kappa (1971) and Gwet's AC1 statistic (2002) are designed to be chance-corrected measures of inter-rater reliability. Correcting for chance is important as these estimates are closer to what one would expect when using the instrument on a different group of ratees (Graham et al., 2012). Table 4.2 shows the percent of overall agreement, Fleiss' kappa (1971), and Gwet's AC1 statistic (2002), along with their standard errors, respectively, for all items together and for each standard assessing teacher quality individually when combining the data from all four videos.

Table 4.2
Inter-rater Reliability for All Four Videos Combined

	Fleiss' Kappa	S.E.	Gwet's AC1	S.E.	Percent Agreement	S.E.
All Items	0.335	0.027	0.595	0.036	0.690	0.022
Standard 1	0.176	0.034	0.294	0.071	0.506	0.039
Standard 2	0.400	0.044	0.661	0.053	0.735	0.034
Standard 3	0.241	0.060	0.570	0.080	0.665	0.050
Standard 4	0.511	0.099	0.776	0.067	0.818	0.047
Standard 5	0.230	0.085	0.700	0.124	0.749	0.086

N = 162

The inter-rater reliability coefficients for All Items collectively are 0.335, 0.595, and 0.69 for Fleiss' kappa, Gwet's AC1, and the percent of overall agreement respectively. There is very little consensus on what signifies acceptable values of inter-rater reliability coefficients. The percent of overall agreement value of 0.69 signifies a moderate level of agreement, though Stemler (2004) suggests 0.75 as signifying an acceptably high value of the percent of overall agreement. The value for Fleiss' kappa (1971) is very low, and does not meet the previously identified acceptable high agreement threshold of 0.75 (Fleiss, 1981), though the value of Gwet's AC1 is much more acceptable. These findings offer some support for the moderate reliability of the observation instrument.

The standard errors of measurement, when examining the coefficients for all videos combined (see Table 4.3), are relatively small when combining all items together (0.022 to 0.036). These values increase when examining the standards individually. Care should be taken when using the coefficients for individual standards specifically as there are less data informing the findings.

As both Gwet's AC1 (2002) and Fleiss' kappa (1971) are chance corrected coefficients, one would expect, should the possible paradox not be evident in this teacher

evaluation context, to be very similar in magnitude. In examining the statistics for Fleiss' kappa (1971), it is evident from the low calculated values when compared to Gwet's AC1 (2002) statistic, that Fleiss' kappa (1971) does in fact fall victim to the possible paradox previously discussed. When examining *All Items* together, there is a large disparity when comparing Fleiss' kappa (1971) to the other measures of inter-rater reliability. To further illustrate, when examining Standard 5 as an example, the percent agreement for Standard 5 is the second highest of any of the individual standards at 74.9% agreement while at the same time, the value of Fleiss' kappa is the second lowest at .230. Gwet's AC1 (2002), on the other hand, provides a more stable estimate of 0.70. Other standards suffer similar problems but are not as illustrative as is Standard 5.

These same data are plotted in Figure 4.1. The values for Fleiss' kappa are lower than either of the other two measures of inter-rater reliability.

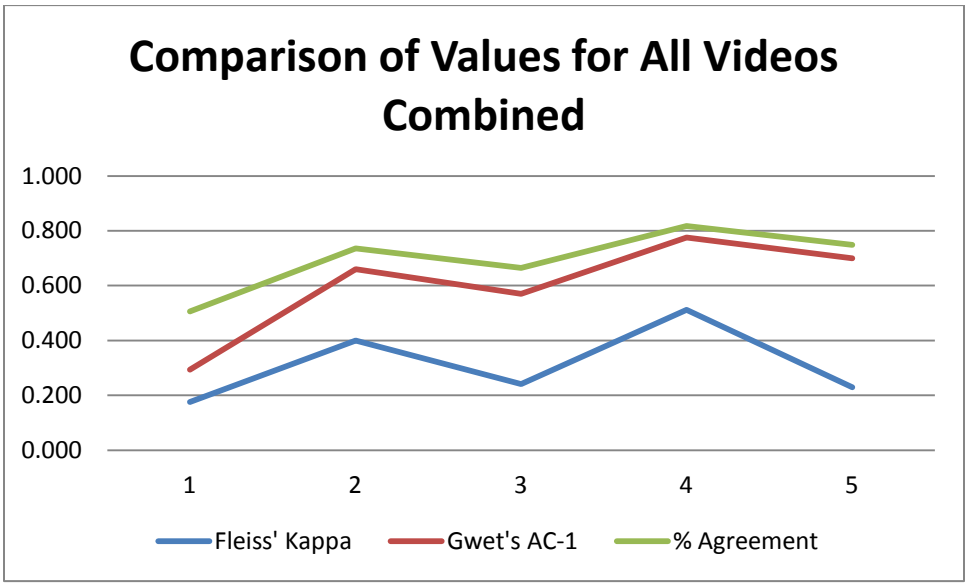

Figure 4.1 Visual Representation of Inter-rater Reliability for All Videos Combined by Standard

While displaying evidence of the paradox which can impact kappa statistics is an important step in this study, studying the behavior of Gwet's AC1 (2002) statistic in inter-rater reliability studies in an educational context is of similar importance (see Research Question 4). As can be seen in Figure 4.1, the values for Gwet's AC1 statistic (2002) are very similar to the percent of overall agreement and follow along a very similar path while the performance of Fleiss' kappa is much more suspect.

A further question influencing this study is to determine if the paradox which often impacts kappa statistics varies when rating teachers which are highly or poorly rated (see Research Question 3). Table 4.3 shows the percent of overall agreement, Fleiss' kappa (1971), and Gwet's AC1 statistic (2002), and their standard errors, respectively, for all items together and for each standard for the highest rated teacher.

Table 4.3
Inter-rater Reliability – Best Rated Teacher

	Fleiss' Kappa	S.E.	Gwet's AC1	S.E.	Percent Agreement	S.E.
All Items	0.075	0.017	0.825	0.039	0.840	0.033
Standard 1	0.011	0.014	0.583	0.095	0.655	0.065
Standard 2	0.086	0.042	0.888	0.053	0.894	0.047
Standard 3	0.057	0.025	0.815	0.090	0.831	0.075
Standard 4	0.018	0.015	0.937	0.039	0.939	0.037
Standard 5	0.031	0.054	0.878	0.129	0.885	0.115

N = 42

There is no clearer illustration, in these data, of the paradox which can impact kappa statistics than when examining the data from the highest rated teacher. As Gwet (2012) stated, "kappa often yields coefficients that are unexpectedly low if compared to the overall percent agreement" (p. 36). Table 4.3 illustrates this phenomenon. The percent of overall agreement for this teacher ranged from 0.655 to 0.939, with an

agreement value for all items combined of 0.840. When comparing these results to the values of Fleiss' kappa (1971), which range from 0.011 to 0.086, the paradox is clear.

The agreement value for all items of 0.84, along with the Gwet's AC1 (2002) statistic of 0.825, suggests that the observation instrument is successful at identifying high-quality teaching and that the principals and assistant principals in the study have similar ideas about what classifies as high-quality teaching.

Further, in investigating the properties of Gwet's AC1 statistic (2002), Figure 4.2 illustrates how Gwet's AC1 statistic (2002) is a much more comparable statistic to the percent of agreement than is Fleiss' kappa (1971), as the patterns across standards are almost identical. In fact, for very high performing teachers, this study suggests that perhaps only Gwet's AC1 statistic (2002) and the percent agreement should be presented when assessing inter-rater reliability when the percent of overall agreement is especially high.

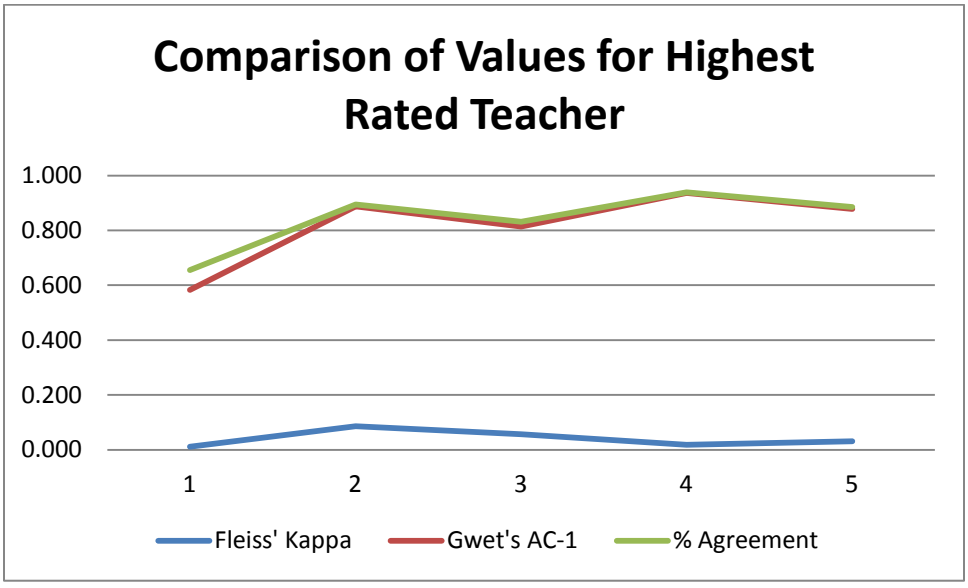

Figure 4.2 Visual Representation of Inter-rater Reliability for Best Rated Teacher by Standard

Finally, just as an investigation of the highest rated teacher provides value to this study, so too does an investigation of the lowest rated teacher add value to this study. Table 4.4 displays the statistics for the lowest rated of the teachers in the study. An examination of how Gwet's AC1 statistic (2002) and Fleiss' kappa (1971) perform when compared to percent agreement still shows support for the use of Gwet's AC1 statistic (2002) even for poorly performing teachers.

Table 4.4
Inter-rater Reliability – Worst Rated Teacher

	Fleiss' Kappa	S.E.	Gwet's AC1	S.E.	Percent Agreement	S.E.
All Items	0.158	0.053	0.334	0.052	0.523	0.032
Standard 1	0.007	0.015	0.157	0.072	0.408	0.037
Standard 2	0.221	0.146	0.477	0.065	0.609	0.050
Standard 3	0.126	0.055	0.247	0.074	0.474	0.045
Standard 4	0.081	0.058	0.491	0.180	0.601	0.114
Standard 5	0.050	0.007	0.093	0.129	0.386	0.061

N = 42

The low inter-rater reliability coefficients associated with the worst-rated teacher in the study suggest that the instrument may be lacking in terms of its ability to successfully identify poor-quality teaching or that what qualifies as poor-quality teaching varies significantly for the principals and assistant principals participating in the study.

While the overall percent of agreement for the lowest rated teacher is much lower across the standards than it was for the highest rated teacher, the values for Fleiss' kappa are all still very low and very different from the percent of overall agreement. Likewise, the values for Gwet's AC1 statistic (2002) are also lower than the percent of overall agreement, though they are much more comparable than are the values of Fleiss' kappa (1971).

The investigation into these inter-rater reliability coefficients suggests that Gwet's AC1 statistic as a measure of inter-rater reliability in educational contexts seems to provide a more stable estimate than does Fleiss' kappa (1971) from these data. It is suggested that Gwet's AC1 statistic (2002) should be included in inter-rater reliability studies in an educational context.

Finally, these data show that the percent of overall agreement is the lowest for the poorest rated teacher. This suggests the possibility that perhaps there is not a very clear picture of what "bad" teaching looks like. This will be expanded on further in Chapter 5.

Figure 4.3 shows visually, once again, how Gwet's AC1 statistic (2002) outperforms Fleiss' kappa (1971) in terms of relation to the percent of overall agreement and suggests that Gwet's AC1 statistic (2002) is a more robust chance-corrected estimate of inter-rater reliability.

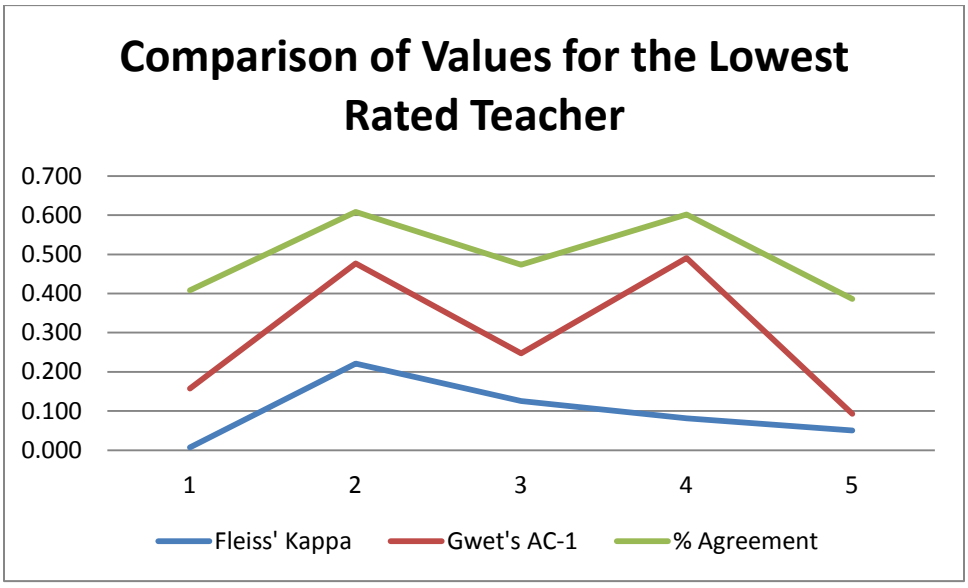

Figure 4.3 Visual Representation of Inter-rater Reliability for Worst Rated Teacher by Standard

Again, while there is some variation in the values, the pattern of the values of Gwet's AC1 statistic is much more similar to that of the percent of overall agreement than is the pattern for Fleiss' kappa, suggesting that Gwet's AC1 statistic should be presented in inter-rater reliability analyses in educational contexts for low performing teachers, as was also found when examining all teachers together and the highest rated teacher alone.

This study was meant to gather and interpret evidence to validate the observational instrument. In doing so, an outline for the steps to follow to validate observation instruments in educational contexts was created. Additionally, Gwet's AC1 statistic was investigated in an educational context to determine if this measure of inter-rater reliability should be included in studies similar to the one conducted here. Evidence suggests that Gwet's AC1 statistic is a robust measure of inter-rater reliability, one that does not suffer from a similar paradox that can impact variations of kappa, and should be included in analyses such as this. These findings have both practical and educational policy implications, and these will be elaborated on in the next chapter.

CHAPTER 5

DISCUSSION OF THE FINDINGS

The purpose of this chapter is to present an overview of the study, a summary of the findings, the principle findings, a discussion of the findings, implications for policy and practice, and outline possible avenues for future research.

Overview of the Study

The purpose of this research is two-fold. One, validation and reliability procedures were outlined in terms of a teacher evaluation context so that observation instruments can be rigorously assessed for quality. The second purpose of this study was to determine if the often reported measures of inter-rater reliability are adequate when assessing inter-rater reliability in an educational context with multiple raters. The typical measures of inter-rater reliability reported in educational research often include a variation of Cohen's kappa (1971) statistic and the percent of overall agreement. Further, an additional measure, Gwet's AC1 statistic (2002), which has been shown in other contexts to be a more stable measure of inter-rater reliability than is Cohen's kappa (1968), was evaluated to determine if this is the case in a teacher evaluation context and to justify its possible inclusion in future reliability studies.

Reliability in a teacher evaluation context is becoming a more important topic of study with the increased emphasis of teacher quality in the Race to the Top federal grants program. This study adds to the literature on the assessment of inter-rater reliability in educational contexts by, one, providing evidence of the paradox from which kappa

statistics can suffer, and two, by illustrating the need to include a more stable inter-rater reliability coefficient when reporting results from an educational reliability study.

The instrument examined was developed in concert by the Clarke County School District and a teacher supervision and classroom observation expert at the University of Georgia. The instrument purported to assess teacher quality and did so by assessing 5 standards comprised of a total of 28 elements.

The data used in the study were gathered from 41 principals and assistant principals within the district using the rating instrument to assess four videotaped lessons from actual teachers within the same school district. These four lessons covered various subjects as well as included two from elementary school teachers, one from a middle school teacher, and one from a high school teacher.

This study resulted in four primary findings. They are discussed in the following section.

Discussion of the Findings

Finding 1: Evidence supporting the content validity of teacher evaluation observation instruments is primarily procedural in nature, and should be well-documented throughout instrument creation.

The observation instrument used in the Clarke County School District to observe and evaluate teacher quality was created in a multistep way, and with the influence of a great deal of expert opinion. The documentation provided by the district included timelines, procedures, personnel involved, training materials, and reference material (including peer-reviewed research) used in instrument creation. As this study was envisioned post-creation of the instrument, and had the primary purpose of establishing

inter-rater reliability of the instrument, the procedural documentation provided was vital to establishing content validity.

Districts which embark on such an undertaking should be mindful of each of the steps which leads to the creation of an evaluation instrument and should take measures to well-document these. The types of evidence needed for validation vary based on the intended uses of the instrument (Kane, 2013). The evidence needed for the content validation of the observation instrument, had the procedural steps from the school district not been so clear and well-documented, would have been difficult to interpret, thus making validation difficult to establish.

Finding 2: The paradox which can impact variations of the Cohen's kappa statistic to assess inter-rater reliability is evident in this teacher evaluation context. Additionally, this paradox is not only visible when examining all teachers combined, but can be seen when examining the highest and lowest rated teachers individually.

As expressed in the discussion of related literature in Chapter 2, Cohen's kappa (1968) and variations such as Fleiss' kappa (1971) can suffer from a paradox that can result in artificially low values of kappa when compared to the percent of overall agreement (Feinstein & Cicchetti, 1990; Graham et al., 2012; Gwet, 2002, 2012). Using this teacher evaluation context, evidence of this paradox is evident.

When examining the results of this study for all four teacher videos combined, as seen in Table 4.2, it is fairly clear that there is a large disparity when comparing Fleiss' kappa (1971) to the other measures of inter-rater reliability. To be specific, when examining Standard 5 as an example, the percent agreement for Standard 5 is the second

highest of any of the individual standards at 74.9% agreement while at the same time, the value of Fleiss' kappa is the second lowest at 0.230.

Additionally, when examining measures of inter-rater reliability for the highest and lowest rated teachers individually, similar results are found. These results are perhaps best examined visually. Figure 4.2 and Figure 4.3 show the comparison of the values across each standard for the highest rated and lowest rated teacher participating in the study. The comparison of Fleiss' kappa to the percent of overall agreement shows dramatic differences supporting the notion that the paradox suffered by kappa statistics exists in when examining both highly and poorly rated teachers.

Finding 3: Gwet's AC1 statistic appears, from these data, to be a more robust estimator of inter-rater reliability than is the kappa statistic when compared to the percent of overall agreement, and should be reported in educational contexts. Additionally, this finding holds when examining the highest and lowest rated teachers individually, but seems especially justified for use with highest rated teachers.

Using Table 4.2, it is fairly easy to determine that the values for Gwet's AC1 statistic are much more similar in magnitude to the percent of overall agreement than are the values for Fleiss' kappa. Gwet's AC1 statistic seems to be more robust as an estimate of inter-rater reliability and these findings seem to support the use of Gwet's AC1 statistic in educational evaluation contexts when calculating inter-rater reliability.

As can be seen in Figure 4.2 and Figure 4.3, the pattern and magnitude of the values for Gwet's AC1 statistic is more similar to the pattern and magnitude of the values for the percent of overall agreement than are the values for Fleiss' kappa for both the highest and lowest rated teacher. This finding is particularly true for the highest rated

teacher, which is expected because highly rated teachers have a higher percentage of overall agreement.

Gwet's AC1 statistic (2002) has been shown, in this study, to be a more accurate measure of inter-rater reliability than is the often report kappa statistic when compared to the percent of overall agreement and should be presented in studies such as this. In fact, Gwet's AC1 statistic (2002) outperformed the kappa statistic when using data from all four teachers combined, when only using the highest rated teacher, and when only using the lowest rated teacher, providing support to report this statistic regardless of the educational rating data under review.

Finding 4: The percent of overall agreement is lowest for the lowest rated teacher, suggesting that there may be a gap in the ability to agree what poorer teacher quality looks like.

The value for the percent of overall agreement, when examining all of the teachers combined, is 0.69 for all standards combined, and ranges from 0.51-0.82 when examined by standard. When examining the highest and lowest rated teacher individually, the value for the highest rated teacher for all standards combined is 0.84, with a standard range of 0.66-0.94, while the lowest rated teacher has a value of 0.52 for all standards combined and a range of 0.39-0.61 for individual standards.

These findings suggest that raters may have a much keener sense of what represents high quality teaching, but a much less developed idea of what represents lower quality teaching. Seemingly, the raters agree when observing a high quality teacher but have ratings with much greater variation when observing a poorer quality teacher.

Implications for Policy

With the broad reach of Race to the Top (2009) on teacher evaluation, the findings in this study could have fairly broad impact on the future evaluation of the instruments used to assess teacher quality. As was documented in Chapter 2, the typically reported measures of inter-rater reliability in educational contexts with multiple raters are the percent of overall agreement and some variation of Cohen's kappa (1968). The findings in this study suggest, at a minimum, that Gwet's AC1 (2002) statistic should be included in the assessment of inter-rater reliability for observation instruments such as the one used in this study, and should perhaps be used in lieu of the kappa statistic in similar situations.

Additionally, the finding that there is greater variation in the ratings of the lowest rated teacher suggests that policies may need to be developed which assist in training raters to better identify poorer teacher quality. The standards and elements on the observation instrument, and the training which helped raters accurately identify teacher quality; seem to perform admirably when examining high quality teachers. The findings suggest that something about the standards and elements, or the training, fails to address teacher quality on the poorer end of the spectrum, which could impact both policy and could provide an avenue for future research.

Implications for Practice

This study provides encouraging results to the school district which designed and implemented the observation instrument under review. The overall performance of the instrument, with regard to inter-rater reliability, performed admirably, though with some exception. Future use of the instrument seems justified, though some revision, either to

the standards and elements, or the training, is recommended to better identify teachers at the lower end of the teacher quality spectrum.

Additionally, the procedures the district used to assure content validity prior to the assessment of inter-rater reliability completed in this study can have broad-ranging impact as other states and districts undertake a similar process.

Avenues for Future Research

This quantitative study is meant to serve as a basis for similar studies to draw from. The inter-rater reliability argument presented in favor of the use of Gwet's AC1 (2002) statistic in similar studies is meant to enhance the information typically provided. Additionally, this study informs about the procedures used to insure content validity in a teacher evaluation context. Based on the findings of this study, the following avenues of future research are presented.

1. Qualitative studies which can suggest additional standards and elements which may be able to increase the precision of measurement at the lower end of the teacher quality spectrum should be undertaken. Then a similar study to this one should be conducted again to determine if these new standards and elements had the desired effect.
2. While not of interest to the current study, teacher evaluation scores should be correlated to variables such as student achievement measures to provide additional validity evidence to further validate the observation instrument used in this teacher evaluation system.
3. A qualitative study interviewing the raters in this study focusing on the train of thought followed when rating the lowest performing teacher should be

completed. This study could have an impact on both the future training of the raters and can provide insight into how to increase the precision of the estimate at the lower end of the teacher quality spectrum.

4. A study assessing the internal consistency of the items on the instruments should be conducted to provide further validity evidence about the observation instrument.

Concluding Thoughts

Ultimately, the primary contribution of this study to the field of evaluation lies in the finding that this study provided some evidence suggesting that the lower end of the teacher quality spectrum is a place where extensive research is warranted. While the instrument performed admirably for the best rated teacher, the lack of precision for teachers at the lower end of the spectrum could lead to poor teachers "slipping through the cracks" by being improperly rated. The ultimate goal of any teacher evaluation system is to both accurately identify teacher quality and improve teaching, and in turn, learning and student achievement. A critical avenue to assist in achieving this goal, as it seems from this study, is to improve the precision with which poorer teacher are rated. Future research should be targeted to this aim.

REFERENCES

Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational Researcher*, 37(2), 65-75. doi: 10.3102/0013189X08316420.

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. doi: 0734-306X/2007/2501-0004.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC. American Educational Research Association.

Barge, J.D.. (2012). *School keys: Unlocking excellence through the Georgia school standards*. Atlanta, GA Georgia Department of Education. Retrieved from: http://www.gadoe.org/School-Improvement/Documents/SCHOOL%20KEYS%20FINAL%203-13-12.pdf.

Bolton, D. L. (1972). *Selection and evaluation of teachers*. Berkley:McCutchen.

Buddin, R., & Zamarro, G. (2009). Teacher qualifications and student achievement in urban elementary schools. *Journal of Urban Economics*, *66*(2), 103-115. Retrieved from: http://www.journals.elsevier.com/journal-of-urban-economics/.

Clifford, M. & Ross, S. (2011). Designing principal evaluation: Research to guide decision-making. Washington , D.C.: National Association of Elementary School Principals.

Cohen, J. (1968). "Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit". *Psychological Bulletin,* 70(4), 213–220. doi: 10.1037/h0026256.

Cook, D.A., & Beckman, T.J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine*, 119, 166.e7-166.e16. doi: 10.1016/j.amjmed.2005.10.036.

Corcoran, S. P. (2010). *Can teachers be evaluated by their students' test scores? Should they be? The use of value-added measures of teacher effectiveness in policy and practice.* Providence, Rhode Island: Annenberg Institute for School Reform at Brown University. Retrieved from: http://www.annenberginstitute.org/pdf/valueaddedreport.pdf.

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

Danielson, C. (2011). Evaluations that help teachers learn. *Educational Leadership*, 68(4), 35-40. Retrieved from: http://www.ascd.org/publications/educational-leadership/archived-issues.aspx.

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

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

Darling-Hammond, L. & Millman, J. (1990). The new handbook of teacher evaluation. Newbury Park, CA: Sage Publications, 1990.

Downing, S.M., & Haladyna, T.M. (2004). Validity threats: Overcoming interference with proposed interpretations of assessment data. *Medical Education*, 38, 327-333. doi: 10.1046/j.1365-2923.2004.01777.x.

Ellett, C. D. (1980). Assessing minimum competencies of beginning teachers: Instrumentation, measurement issues and legal concerns. *Evaluation of teaching: The formative process*. *Hot topics series*. Bloomington, IN: Phi Delta Kappa.

Ellett, C. D. (1987). Emerging teacher performance assessment practices: Implications for the instructional supervision role of school principals. In W. Geenfield (ed.), *Instructional leadership: Concepts, issues, and controversies*. Boston: Allyn & Bacon.

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

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of clinical epidemiology*, *43*(6), 551-558. doi: 10.1016/0895-4356(90)90159-M.

Fenstermacher, G. D., & Richardson, V. (2005). On making determinations of quality in teaching. *Teachers College Record*, 107(1), 186–213.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378. doi: 10.1037/h0031619.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. 2nd ed. New York: John Wiley

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

Goe, L., & Holdheide, L. (2011). Measuring teachers' contributions to student learning growth for nontested grades and subjects: Research & policy brief. Washington, D.C.: National Comprehensive Center for Teacher Quality. Retrieved from: http://www.eric.ed.gov/PDFS/ED520722.pdf.

Goe, L., Holdheide, L., & Miller, T. (2011). A practical guide to designing comprehensive teacher evaluation systems: A tool to assist in the development of teacher evaluation systems. Washington, D.C.: National Comprehensive Center for Teacher Quality. Retrieved from: http://www.eric.ed.gov/PDFS/ED520828.pdf.

Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. Center for Educator Compensation Reform. Retrieved from: http://cecr.ed.gov/pdfs/Inter_Rater.pdf.

Gwet, K. (2002). Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical Methods for Inter-rater Reliability Assessment*, *1*(6), 1-6. Retrieved from: http://www.agreestat.com/research_papers/kappa_statistic_is_not_satisfactory.pdf

Gwet, K. L. (2012). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among multiple raters*. Gaithersburg, MD: Advanced Analytics, LLC.

Haefele, D. L. (1993). Evaluating teachers: A call for change. *Journal of Personnel Evaluation in Education*. 7(1), 21-31. doi: 10.1007/BF00972346.

Hanover Research. (2011). *A survey of race to the top teacher evaluation systems*. Retrieved from: http://www.hanoverresearch.com/toolkit/pdf/A%20Survey%20of%20Race%20to%20the%20Top%20Teacher%20Evaluation%20Systems%20-%20Membership.pdf.

Hershberg, T. (2005). Value-added assessment and systematic reform: A response to the challenge of human capital development. *Phi Delta Kappan*. 87(4), 276-283. Retrieved from: http://pdkintl.org/publications/kappan/.

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

Jaeger, R. M., & Tittle, C. K. (1980). *Minimum competency testing: Motives, models, measures, and consequences*. Berkeley, CA: McCutchan.

Joint Committee on Standards for Educational Evaluation. (1988). *The personnel evaluation standards: How to assess systems for evaluating educators*. D. Stufflebeam (ed.). Newbury Park, CA: Corwin Press.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*. 112(3), 527-535. doi: 10.1037/00332909.112.3.527.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.

Kane, M. (2009). Validating the interpretations and uses of test scores. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications*. Charlotte, NC: Information Age Publishing, Inc.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*. 50(1), 1-73. doi: 10.1111/jedm.12000.

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

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

Koedel, C., & Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Policy*, 6(1), 18-42. doi: 10.1162/EDFP_a_00027.

Koretz, D. (2008). A measured approach: Value added models are a promising improvement, but no one measure can evaluate teacher performance. *American Educator*, 32(3), 18-27. Retrieved from: https://www.aft.org/pdfs/americaneducator/fall2008/koretz.pdf.

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. doi: 10.2307/3699496.

Linn, R. L., & Baker, E. L. (1996). Can performance-based student assessments be psychometrically sound? In J. B. Baron and D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities, Ninety-fifth Yearbook of the National Society for the Study of Education.* Chicago: University of Chicago Press.

Linn, R. L. (2008). Educational accountability systems. In K.E. Ryan and L.A. Shepard (Eds.), *The future of test-based accountability*. New York: Routledge.

Lohman, J. (2010). *Comparing no child left behind and race to the top* (2010-R-0235). OLR Research Report. Retrieved from: http://www.cga.ct.gov/2010/rpt/2010-R-0235.htm.

Loup, K.S., Garland, J.S., Ellet, C.D., & Rugutt, J.K. (1996). Ten years later: Findings from a replication of a study of teacher evaluation practices in our 100 largest school districts. J*ournal of Personnel Evaluation in Education,* 10(3), 203-226. doi: 10.1007/BF00124986.

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

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

Marzano, R. J. (2012). The two purposes of teacher evaluation. *Educational Leadership*, *70*(3), 14–19. Retrieved from: http://www.ascd.org/publications/educational-leadership/nov12/vol70/num03/The-Two-Purposes-of-Teacher-Evaluation.aspx#interview.

McCaffrey, D. F., & Lockwood, J. R. (2011). Missing data in value-added modeling of teacher effects. *Annals of Applied Statistics*, 5(2A), 773-797. doi: 10.1214/10-AOAS405.

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

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy*, 4(4), 572-606. doi: 10.1162/edfp.2009.4.4.572.

McLaughlin, M., & Pfeifer, R. S. (1988). *Teacher evaluation, improvement, accountability, and effective learning*. New York: Teachers College Press.

McColskey, W., & Egelson, P. (1993). Designing Teacher Evaluation Systems that Support Professional Growth. Office of Educational Research and Improvement. Retrieved from: http://files.eric.ed.gov/fulltext/ED408287.pdf.

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

Mead, S., Rotherham, A., & Brown, R. (2012). The Hangover: Thinking about the Unintended Consequences of the Nation's Teacher Evaluation Binge. Teacher Quality 2.0. Special Report 2. *American Enterprise Institute for Public Policy Research*. Retrieved from: http://www.aei.org/paper/education/k-12/teacher-policies/the-hangover-thinking-about-the-unintended-consequences-of-the-nations-teacher-evaluation-binge/.

Medley, D. M., & Coker, H. (1987). The accuracy of principals' judgments of teacher performance. *Journal of Educational Research*, 80(4), 242-247. doi: 10.2307/40539630.

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

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18(2), 5-11. doi: 10.2307/1175249.

Messick, S. (1995). Validity of psychological assessment. Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749. doi: 10.1037/0003-066X.50.9.741.

Messick, S. J. (Ed.). (1999). Assessment in higher education: Issues of access, quality, student development, and public policy. Mahwah, NJ: Lawrence Erlbaum Associates.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, D.C. Retrieved from: http://datacenter.spps.org/uploads/SOTW_A_Nation_at_Risk_1983.pdf.

National Council of Teachers of English. (2012). *NCTE position statement on teacher evaluation.* Retrieved from: http://www.ncte.org/positions/statements/teacherevaluation.

National Council of Teachers of Mathematics. (2011). *Teacher evaluation: A position of the National Council of Teachers of Mathematics*. Retrieved from: http://www.nctm.org/uploadedFiles/About_NCTM/Position_Statements/Teacher%20Evaluation%20(with%20references,%202011).pdf.

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: Exploration of stability across models and contexts. *Educational Policy Analysis Archives*, 18(23), 1-22. Retrieved from: http://epaa.asu.edu/ojs/article/view/810/858.

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. doi: 10.3102/0002831210362589.

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

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

Raudenbush, S. W., & Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, 20(4), 307-335. doi: 10.2307/1165304.

Recesso, A.M. & Zepeda, S.J. (2008). Evidential reasoning and decision support in assessment of teacher practice. In T.J. Kowalski & T.J. Lasley (eds). Handbook on data-based decision making in education (pp.363-381). Erlbaum.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review*, *94*(2), 247-252. doi: 10.1257/0002828041302244.

Rubin, D. B., Stewart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, 29(1), 103-116. doi: 10.2307/3701308.

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

Sass, T. R. (2008). The stability of value added measures of teacher quality and implications for teacher compensation policy. Policy Brief 4. Washington, D.C.: The Urban Institute, National Center for Analysis of Longitudinal Data in Education Research. Retrieved from: http://files.eric.ed.gov/fulltext/ED508273.pdf.

Scriven, M. (1988). Duty-based teacher evaluation. *Journal of Personnel Evaluation in Education*, 12(3), 247-257. doi: 10.1007/BF00124098.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9(4). http://pareonline.net/getvn.asp?v=9&n=4.

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

Stronge, J. H. (2006). *Evaluating teaching: A guide to current thinking and best practice*, (2nd ed.). Thousand Oaks, CA: Corwin Press.

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

Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M., Roth, J., Ariet, M., Fisher, T., & Resnick, M. B. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics*, 29(1), 11-36. doi: 10.2307/3701305.

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

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

U.S. Department of Education. (2009). *Race to the top executive summary*. Washington, D. C.: Retrieved from: http://www2.ed.gov/programs/racetothetop/executive-summary.pdf.

Waters, J.T., Marzano, R.J., & McNulty, B. (2003). *Balanced leadership: What 30 years of research tells us about the effect of leadership on student achievement*. Aurora, CO: Mid-continent Research for Education and Learning.

Wiener, R., & Jacobs, A. (2011). Designing and implementing teacher performance management systems: Pitfalls and possibilities. Queenstown, MD: The Aspen Institute. Retrieved from: http://www.aspendrl.org/portal/browse/DocumentDetail?documentId=580&download.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). The widget effect: our national failure to acknowledge and act on teacher differences. *Brooklyn, NY: The New Teacher Project.* Retrieved from: http://gcpstv.org/gcps-mainweb01.nsf/092DF14366B4598F8525788C0067CF48/$file/TNTPTheWidgetEffect.pdf.

Wenglinsky, H. (2002). How schools matter: The link between classroom practices and student academic performance. *Educational Policy Analysis Archives*, 10(12). Retrieved from: http://epaa.asu.edu/ojs/article/view/291/417.

Winters, M. A., & Cowen, J. M. (2013). Who would stay, who would be dismissed? An empirical considerations of value-added teacher retention policies. *Educational Researcher*, 42(6), 330-337. doi: 10.3102/0013189X13496145.

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

Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K. (2013). A comparison of Cohen's kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. *BMC Medical Research Methodology*, 13(61), 1-7. doi: 10.1186/1471-2288-13-61.

Zepeda, S. J. (2006). High stakes supervision: We must do more. *International Journal of Leadership in Education*, *9*(1), 61-73. doi: 10.1080/13603120500448154.

Zepeda, S. J. (2012). Instructional supervision: Applying tools and concepts. Larchmont, NY: Eye on Education.

Zepeda, S. J. (2013). The principal as instructional leader: A handbook for supervisors (3rd ed.). Larchmont, NY: Eye on Education.

Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. W. Lissitz (Ed.), The concept of validity: Revisions, new directions and applications (pp. 65–82). Charlotte, NC: IAP-Information Age Publishing, Inc.

Appendix A. Protocol for Validation Study

Clarke County School District Classroom Observation Validation Study

Good morning and thank you for participating in this study.

This study will hopefully provide validation for the Classroom Observation Instrument and strengthen our collaborative work on the CCSD Teacher Evaluation System. For this study to be meaningful, adherence to the validation process protocols is vital.

This morning, you will watch four videos of CCSD teachers teaching a lesson.

There are two elementary teachers, one middle school teacher, and one high school teacher.

Before we begin each video, we will provide you with either a Tier I or a Tier II Classroom Observation Instrument. We will then start the video.

Please turn your cell phones off completely and refrain from using any technology while the validation process is under way.

Once the video begins, we ask that you do not

1. talk or confer about what you are seeing; or,
2. share the rating instruments with each other.

Please treat this as if you were physically observing instruction in the classroom, and use the Classroom Observation Instrument to score instruction accordingly.

At the end of the video, you will be given 10 minutes to complete your Classroom Observation Instrument and provide comments in each Performance Standard Section.

After 10 minutes is up, the R&D team will collect the forms and prepare for the next video. This process will repeat four times, and you will then be released for the afternoon.

Even after the study has concluded, please do not discuss what you had seen or how you scored any of the standards and their elements.

Your assistant principals will participate in the very same process this afternoon, and central office leaders will be involved in the same activity at a later date. That said, please do not share details of your experience with your assistant principals or any central office administrator.

In a validation study, the process and the protocols are everything.

Please assist us to follow these procedures to the letter.

Appendix B. Sample of Observation Instrument Used in Study

Clarke County School District
Tier I Teacher Observation Form

Teacher's Name: _____#1_____ School: _Cleveland Rd_____ Grade: __Elem_ Subject: __Read___

Type of Observation: ☒ Announced ☐ Unannounced Date Observed: __1/17/13____ Start Time: _10:00____

End Time: ___10:33_____ Number of Students Present: _____26_____

Evaluator's Name:_____ Position: Principal/Assoc. or Asst. Principal/Central Office_____

Curriculum and Planning (Standard CP): The teacher consistently makes decisions about planning that demonstrate a deep understanding of grade level content knowledge, pedagogy, and CCGPS or State-approved curriculum implementation by appropriately planning for what students are expected to know, understand, and be able to do.

	CP Standard Elements			
CP 1	The teacher uses the Clarke County School District Instructional Framework to support standards-based instruction.	☐ Yes	☐ No	☐ N/A
CP 2	The teacher uses the required curriculum to plan instruction that reflects strong knowledge and understanding of both content and effective instructional delivery.	☐ Yes	☐ No	☐ N/A
CP 3	The teacher plans assessments to measure student progress toward and mastery of the curriculum.	☐ Yes	☐ No	☐ N/A
CP 4	The teacher plans for appropriate differentiation.	☐ Yes	☐ No	☐ N/A
CP 5	The teacher plans instruction that includes cross-curricular and/or global connections.	☐ Yes	☐ No	☐ N/A

CP Comments:

Standards-Based Instruction Standard (SBI): The teacher consistently uses research-based practices in the classroom, challenging all learners to achieve high levels of learning as defined by CCGPS or State-approved curriculum.

	SBI Standard Elements			
SBI 1	The teacher sequences the lesson or instruction in a logical, predictable manner that includes an opening, mini-lesson, work session, and closing.	☐ Yes	☐ No	☐ N/A
SBI 2	The teacher effectively communicates learning expectations by using the language of the standards.	☐ Yes	☐ No	☐ N/A
SBI 3	The teacher uses a variety of instructional strategies and delivery modes which are used to address student needs.	☐ Yes	☐ No	☐ N/A
SBI 4	The teacher demonstrates the use of practices that engage students in learning.	☐ Yes	☐ No	☐ N/A
SBI 5	The teacher demonstrates high expectations for all students.	☐ Yes	☐ No	☐ N/A
SBI 6	The teacher implements/incorporates technology to enhance instruction.	☐ Yes	☐ No	☐ N/A
SBI 7	The teacher differentiates instruction based on student needs.	☐ Yes	☐ No	☐ N/A
SBI 8	The teacher delivers instruction which develops higher order thinking skills.	☐ Yes	☐ No	☐ N/A
SBI 9	The teacher delivers instruction to include cross-curricular and/or global connections.	☐ Yes	☐ No	☐ N/A
SBI Comments:				
Assessment of Student Learning Standard (AL): The teacher consistently uses a balanced variety of assessment techniques that are systematically implemented, resulting in appropriate interventions that foster continuous improvement for all.				
	AL Standard Elements			
AL 1	The teacher intentionally solicits feedback from students on their understanding of the standard.	☐ Yes	☐ No	☐ N/A
AL 2	The teacher uses assessment tools/strategies to diagnose individual and class strengths, misconceptions, and areas of weaknesses.	☐ Yes	☐ No	☐ N/A

AL 3	The teacher uses formative assessment tools/strategies to identify individual class strengths, misconceptions, and areas of weaknesses.	☐ Yes	☐ No	☐ N/A
AL 4	The teacher adjusts instruction to maximize student achievement of the CCGPS or state approved curriculum.	☐ Yes	☐ No	☐ N/A
AL 5	The teacher provides effective and timely standards-based verbal/written feedback regarding student work.	☐ Yes	☐ No	☐ N/A
AL Comments:				
Instruction Environment Standard (IE): The teacher consistently creates a safe, productive, collaborative, and inviting learning environment that fosters a sense of community and personal responsibility to ensure that students maximize learning.				
	IE Standard Elements			
IE 1	The teacher establishes classroom rituals and routines that support a positive, productive learning environment.	☐ Yes	☐ No	☐ N/A
IE 2	The teacher maximizes instructional time.	☐ Yes	☐ No	☐ N/A
IE 3	The teacher fosters a sense of community and belonging by acknowledging diversity, achievements, and/or accomplishments of learners in the classroom.	☐ Yes	☐ No	☐ N/A
IE 4	The teacher helps students take responsibility for their own behavior and learning.	☐ Yes	☐ No	☐ N/A
IE Comments:				
Building Positive Student Relationships (BPSR): The teacher consistently strives to develop connections/relationships with students based on the district core values and sees this as a favorable and necessary condition of learning.				
	BPSR Standard Elements			
BPSR 1	The teacher seeks to understand the social and emotional needs of students.	☐ Yes	☐ No	☐ N/A
BPSR 2	The teacher recognizes the importance of building rapport with students and sees this as beneficial to student achievement.	☐ Yes	☐ No	☐ N/A

BPSR Comments:				

1. If you were using ONLY this observation as a way to assess this teacher, which rating would you give this teacher on his/her overall performance? Check One:

___ Exemplary: consistently exceeds
___ Proficient: Frequently meets
___ Emerging: Sometimes meets
___ Unsatisfactory: Does not meet

2. List two or three items you would want to probe this teacher on during the post-observation conference?

Appendix C. Procedural Steps Used in the Creation of the Observation Instrument

Step Item	Procedures	Time Frame
1	There was recognition that the Georgia Teacher Evaluation System (GTEP) was inadequate to assess procedures in the Clarke County School District (CCSD) and a new tier-based evaluation system focused on teacher experience and effectiveness was envisioned.	Fall – 2010
2	Draft I of the CCSD Teacher Evaluation System was created based on the Class Keys and the CCSD Non-Negotiable Practices for High Student Performance. These documents were reviewed by the Superintendent's cabinet.	Spring – 2011
3	A Principal Teacher Evaluation Advisory Committee was assembled to offer feedback on Tier I and Tier II of the CCSD Teacher Evaluation System. The system was pilot-tested with a small group of principals, assistant principals, and teachers. Refinements were made based on the feedback from the pilot.	Spring – 2011

Step Item	Procedures	Time Frame
4	The CCSD/UGA system-wide Professor-in-Residence was enlisted to support the work by further developing and refining the CCSD Teacher Evaluation System. The objective was to anchor best practices associated with transformations within the district but also be based on the translations of research.	Spring – 2011
5	The Professor-in-Residence further refined classroom observation forms, paying special attention to both content, legal issues, translations of research, and processes relevant to teacher evaluation, focusing primarily on classroom observation protocols.	Summer – 2011
6	Two full-day training sessions were held for all school leaders.	Summer – 2011
7	All teachers were formally schooled in the intents, purposes, processes, and procedures of the CCSD Teacher Evaluation System.	Summer – 2011

Step Item	Procedures	Time Frame
9	Changes resulting from refinement procedures were reviewed quarterly by the Principal Teacher Evaluation Advisory Committee.	Fall – 2011
10	Additional documents (e.g., CCSD Classroom Observable Practices) to support the accurate assessment of teacher quality were created for principals and assistant principals.	Fall – 2011
11	On-going training at monthly principal meetings; quarterly meetings of the Principal Teacher Evaluation Advisory Committee. An extended scoring guide for Exemplary, Proficient, Emerging, and Unsatisfactory was developed.	Spring - 2012
12	Development of CCSD Annual Evaluation Performance Rubric; Refinement of Observation Instruments for Tier I and Tier II based on lessons learned from Year I and Year II; continuation of training for principals and assistant principals; vetting of Annual Evaluation Performance Rubric	Summer – 2012

Step Item	Procedures	Time Frame
13	Implementation of Annual Evaluation Performance Rubric; ongoing training of principals and assistant principals; refinement of documents and vetting of materials with Principal Teacher Evaluation Advisory Committee	Fall 2012
14	System-wide self-study on VAM and Artifacts and Evidence related to classroom observation and teacher evaluation system; Crosswalk of state of Georgia Teacher Keys observation instrument and CCSD teacher observation instrument; Development of new categories and observation instrument for Fall 2013	Spring 2013