Impact of Teacher Evaluation Protocols

on Classroom Instructional Practices

by

Kathleen Kwolek

Submitted in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Education

in the

Educational Leadership

Program

YOUNGSTOWN STATE UNIVERSITY

December, 2014

Impact of Teacher Evaluation Protocols

on Classroom Instructional Practices

Kathleen A. Kwolek

I hereby release this dissertation to the public. I understand that this dissertation will be made available from the OhioLINK ETD Center and the Maag Library Circulation Desk for public access. I also authorize the University or other individuals to make copies of this dissertation as needed for scholarly research.

Signature:

Kathleen A. Kwolek, Student Date

Approvals:

Dr. Karen H. Larwin, Dissertation Advisor Date

Dr. Jane Beese, Committee Member Date

Dr. Kenneth Miller, Committee Member Date

Dr. Darwin W. Huey, Committee Member Date

Dr. Salvatore A. Sanders, Associate Dean of Graduate Studies Date

Abstract

Many states are in the process of adopting rigorous, standards-based teacher evaluation systems in order to address and increase teacher accountability for student achievement. In the newly adopted Educator Effectiveness evaluation system, Pennsylvania classroom teachers were evaluated through one of two evaluation protocols, both aligned to Danielson's *Framework for Teaching*. The purpose of this study was to explore the impact of the new Pennsylvania evaluation model on the use of classroom instructional practices by teacher participants.

The study sample included 111 classroom teachers in grades K-12 in one school district: 35 teachers were assigned to the Formal Observation Model and 76 teachers to the Differentiated Supervision Model (which used portfolios to evaluate teacher performance). A two-group, pretest-posttest, quasi-experiment was used to compare the ratings of classroom instructional practices of all teachers. Using a paired-samples *t*-test, the average increase in the ratings of classroom instructional practices of Formal Observation participants was not significant, whereas the average increase in the ratings of Differentiated Supervision participants was significant.

This study provides evidence that teachers' use of a carefully structured portfolio as a reflection tool may result in improved classroom instructional practices; however, the final Educator Effectiveness ratings of teachers in the Portfolio Mode lacked the discrimination necessary to meet the summative goals for teacher evaluation. These results are important considerations for PA district leaders who must choose among various options for the Differentiated Supervision evaluation of teachers.

Keywords: teacher evaluation, differentiated supervision, Educator Effectiveness ratings, classroom instructional practices, portfolios, Danielson's *Framework for Teaching*

Acknowledgements

Without the support and guidance of the YSU faculty, acquiring the knowledge base and research skills required to complete this dissertation would not have been possible. I deeply appreciate the input of my dissertation committee members, Dr. Jane Beese, Dr. Kenneth Miller, and Dr. Darwin Huey, in helping me plan and construct this project and analyze the results and implications of the research findings. However, without Dr. Karen Larwin, my dissertation advisor, teacher, and mentor, I could never have tackled the obstacles inherent in statistical research; her support and encouragement in every step of the process was instrumental to the completion of this dissertation.

To my family, thank you for your encouragement throughout these past years. The time and dedication necessary to complete such a project was only possible by your enduring support. I would also like to thank my educational colleagues and all the teachers in this study – as a result of my observations, I have a much greater understanding and appreciation of the passion for teaching and learning you bring to your classrooms on a daily basis. In some small way, I hope my pursuits in completing this research project inspire others to act on the belief that life-long learning is possible at any stage of your career.

Table of Contents

Chapter 1

Statement of the Problem

Teaching is a complex process (Lavy, 2007; Marshall, 2005; Phillips & Weingarten, 2013; Scherer, 2012; Seidel & Shavelson, 2007), requiring knowledge of subject matter, pedagogy, and classroom management skills. Historically, evaluation was focused on how well teachers performed and demonstrated these skills, resulting in a summative rating with limited potential to impact student learning. With the advent of the accountability movement and focus on student achievement, teachers must also possess the ability to engage, motivate, and differentiate instruction for all students. This shift, from a focus on the teaching to a focus on the learning, has produced major differences of opinion in how teachers should be evaluated. Implementing a teacher evaluation system that addresses the complexity of teaching while helping teachers improve student learning is an important and critical undertaking (Marshall, 2005; Phillips & Weingarten, 2013; Scherer, 2012). Use of teacher evaluation models as a formative assessment of the teaching and learning processes has the potential to improve classroom instruction. The proposed research will investigate the impact of teacher evaluation protocols on classroom instructional practices.

Purpose Statement

Although the primary purpose of evaluating the effectiveness of teachers should focus on the improvement of student learning, summative teacher ratings provide little information for addressing this goal. In the last decade, the accountability movement has brought intense scrutiny and attention to local and state initiatives aimed at improving their processes of teacher evaluation (Alvarez & Anderson-Ketchmark, 2011). As a result of the failure of traditional teacher evaluation systems to distinguish between effective and ineffective teachers (Bill & Melinda Gates Foundation, 2011; Weisberg, Sexton, Mulhern, & Keeling, 2009), states and local school districts are considering major overhauls of their teacher evaluation processes, typically incorporating some type of metric to differentiate between effective and ineffective teaching. There is significant variability and even controversy regarding how the effectiveness should be measured (Kane, Taylor, Tyler, & Wooten, 2011).

Traditional teacher evaluation systems are perfunctory, annual observations, consisting of a laundry list of performances and classroom attributes (look-fors) that result in an overall satisfactory/unsatisfactory designation, discouraging ratings that more accurately reflect the competencies of teacher effectiveness, and providing no indication of strengths or areas for improvement (Kimball, 2002; Marshall, 2005). Using traditional models, nearly 100% of teachers receive satisfactory evaluations in the vast majority of school districts across the nation (Kane et al., 2011; Weisberg et al., 2009) despite the tremendous range of different instructional methods observed in classrooms and resulting impact on student achievement. In traditional evaluation systems, teachers typically operate autonomously within their own classrooms for a vast majority of their daily lessons. Periodically (usually once or twice a year), an administrator enters the classroom and records some notes about the lesson; unless there are notable deficiencies, this process of teacher evaluation has little impact on the quality of instruction occurring in these classrooms.

In 1995, the release of the student achievement results in the Trends in International Mathematics and Science Study (TIMSS) generated national awareness and concern of deficiencies in our existing educational system. Since then, standardized test results have been used to compare states, districts, and schools within our country. Teacher evaluation methods that use student achievement data provide school leaders with summative information regarding the effectiveness of classroom teachers; unfortunately, the data are not available until the end of the school year, when teachers no longer have the students. The use of achievement data to evaluate individual teachers does not take into account the various factors over which teachers have no control: resources, support, teaching loads, students' prior learning, class sizes, and other influences that affect student test performance (Haertel, 1986). In addition, achievement data provide no information to teachers or leaders regarding the success of classroom behaviors or instructional practices.

The latest trend in evaluation protocols is the inclusion of powerful statistical models that measure annual student growth in selected subjects and grade levels (value-added growth models). While the traditional observation tool evaluated the inputs (lesson plan, classroom preparation, teaching methods, classroom assignments, questioning techniques, etc.) provided by teachers during the course of the year, the statistical models evaluate the outputs (growth in achievement of students as measured by standardized test scores). While some may view traditional observations as obsolete, biased, and ineffective for making summative decisions regarding teacher effectiveness, the value-added growth models do not provide educational leaders with information that will reveal the strengths and weaknesses of classroom instruction. Just as critical, value-added measures cannot help educators understand why certain practices or behaviors are more successful than others. Scores do not inform evaluators or teachers with information on improving classroom effectiveness (Measures of Effective Teaching, [MET], 2010). As a result, teachers performing at moderate levels will not get the critical professional development and differentiated support necessary to improve their effectiveness (Weisberg et al., 2009). In non-tested subjects and grade levels, value-added models are not available to base any decisions regarding teacher effectiveness.

Many states are in the process of adopting a rigorous, standards-based teacher evaluation system in order to address and increase teacher accountability for student achievement. Three years of research examined the use of evaluation systems based on the Charlotte Danielson *Framework for Teaching* and documented a positive correlation between teachers' evaluation scores and student achievement (Heneman, Milanowski, Kimball, & Odden, 2006), suggesting that the instructional practices measured by Danielson's framework contribute to student learning (PACER, 2011).

In Pennsylvania, teacher evaluations became the focus of the fall, 2011 legislative session (PACER, 2011). Prior to this, there had only been one change to the Pennsylvania Public School Code of 1949 regulating evaluation of professional educators, affecting only non-tenured staff. The decades-old, one-half page teacher evaluation form did not require more than an annual endorsement by the teacher's supervisor, resulting in either a satisfactory or unsatisfactory rating. After a three-year pilot of the Danielson *Framework for Teaching,* a new evaluation system was adopted on June 21, 2012 (PA Code, 2013), effective for the 2013-14 school year. The system, known as the Educator Effectiveness rating tool, requires 85% of a classroom teacher's annual evaluation be based on their ratings in each of the four domains of the Danielson *Framework for Teaching,* and 15% be based on a new School Performance Profile rating for each school building in the

Commonwealth (these percentages are in place for 2013-14 and will change to include student-specific assessment and value-added data in subsequent years). Districts are directed to cycle classroom teachers through one of two evaluation protocols, both based on the Danielson framework: (1) the Formal Observation Model, consisting of a pre-observation conference with the teacher, a classroom observation rating teachers on each of the four domains of the framework, and a post-observation conference between the teacher and the administrator, or (2) the Differentiated Supervision Model, which consists of an approved alternative to the Formal Observation process. Three modes of Differentiated Supervision have been approved thus far: Peer Coaching, Action Research, or Portfolios. A three to five year cycle is recommended in which teachers participate in the Formal Observation once during this cycle and participate in the Differentiated Supervision Model in remaining years. The school district in this study selected a four-year cycle to meet the mandates of this legislation.

Using evaluations that give teachers accurate feedback on their teaching provides an opportunity for them to reach their full potential (The New Teacher Project [TNTP], 2012). Although feedback alone is not expected to change instruction, it is likely to enhance teacher reflection and growth (Kimball, 2002) and research indicates improved professional dialogue between administrators and teachers using standards-based teacher evaluation models (Heneman, Milanowski, Kimball, & Odden, 2006). Instead of focusing on how ratings obtained from an evaluation system accurately reflect pre-conceived notions of teacher effectiveness, this research investigates areas not addressed in the research: Will the implementation of a standards-based teacher evaluation model have a positive impact on teachers' use of instructional best practices?

Significance

While rating teacher effectiveness has become a national priority and the subject of most contemporary research, the improvement of teaching practices is more likely to result in improved student learning. Regardless of the method chosen to evaluate teacher effectiveness, unless the process results in the continuous use of best practices of classroom instruction by teachers, improvements in student achievement are unlikely to occur. Examination of the potential for an evaluation model to foster sustained use of classroom best practices can provide the educational community with guidance in the implementation of a teacher evaluation system.

Purpose

The purpose of this study is to explore the impact of the new Pennsylvania state-mandated, high-stakes teacher evaluation model on the use of classroom instructional practices by teacher participants. While state educational policymakers across the country are assuming that teacher accountability for student achievement will improve through the adoption of rigorous models of teacher evaluation, the proposed research will investigate the potential of this model to improve instructional practices in classroom instruction.

The following relationships among different measurements of teachers' classroom practices will be explored:

1. Observations of Classroom Instructional Practices, measured by the researcher before and after the teachers' participation in one of two models of the new teacher evaluation protocol: (1) Formal Observation, and (2) Differentiated Supervision;

2. Summative Educator Effectiveness ratings of teachers in each of the evaluation protocols;
3. Comparison of the individual teachers' scores obtained during the final observation of their Classroom Instructional Practices by the researcher with their summative Educator Effectiveness ratings completed by the principal/supervisor; and
4. Teachers' beliefs of Self-Efficacy in specific components of classroom instruction, measured at the beginning and end of the school year.

Hypothesis

Based on a review of the literature, it is hypothesized that there is a relationship between teacher participation in a Formal Observation Model and implementation of best practices in classroom instruction. Specifically, it is hypothesized that teachers participating in a Formal Observation Model will implement changes and improvements in classroom instructional practices to a greater extent than teachers participating in the Differentiated Supervision Model.

Definition of Key Terms

Best practices for Classroom Instruction– for purposes of this study, best practices are teacher-specific instructional practices recommended by educational research to improve student learning.

Differentiated Supervision Model – as described in PA's Act 82 legislation (PA Code, 2013), this is an alternative to Danielson's Formal Observation Model of teacher evaluation. Examples of acceptable modes for this model include, but are not limited to: Peer Coaching, Self-Directed/Action Research, and Portfolios.

Educator Effectiveness Ratings – in this study, these ratings are the summative evaluation scores received by classroom teachers for their 2013-14 annual evaluation. Each teacher will be evaluated by an administrator in each of the four domains of the Danielson Framework for Teaching, using a zero-to-three point rubric.

Formal Observation Model – as described in PA's Act 82 legislation (PA Code, 2013), teacher evaluation will be based on Charlotte Danielson's Framework for Teaching. Every classroom teacher will be evaluated in the Formal Observation Model at least once during a cycle of years to be determined by the Local Education Agency (LEA). The Formal Observation Model consists of a pre-conference, classroom observation, post-conference, and follow-up walk-throughs by an administrator. Collaborative reflections on these observations and other evidence related to one of the four domains of framework will be used by the administrator to generate the teacher's annual evaluation.

Formative Evaluations – provide teachers with feedback on how to improve instructional performance.

Peer Coaching Mode – a sample option for Differentiated Supervision in PA's Act 82 legislation (PA Code, 2013). Professional employees work collaboratively in small groups to discuss and observe each other's pedagogy, student learning, alignment of curriculum, or other professional needs. Documentation to be shared with the principal/supervisor and used as evidence for the evaluation process will include: rationale for selected target goals, plans to address identified areas of need, dates of observations, data collected, and notes from reflective sessions.

Portfolio Mode – a sample option for Differentiated Supervision in PA's Act 82 legislation (PA Code, 2013). Professional employees will examine and reflect on their own practice in relation to the Danielson Framework for Teaching. A written report and/or documented discussions with colleagues will be used by the principal/supervisor as evidence for the evaluation process.

Scree Test – a statistical analysis used to determine the important factors which account for the bulk of the correlations in the matrix.

Scope of Influence – the size of the group over which one has influence; in the context of classrooms, it refers to influence over individual students in the classroom compared to influence over a group of students as a whole.

Self-Directed/Action Mode – a sample option for Differentiated Supervision in PA's Act 82 legislation (PA Code, 2013). In this mode, classroom teachers may work alone or in small groups to complete an action research project. Documentation to be shared with the principal/supervisor and used as evidence for the evaluation process will include meeting notes, resources, data collection tools, and on-going reflections of a practice-related issue.

Standards-Based Teacher Evaluation Models – teacher evaluations that identify and measure the instructional strategies, professional behaviors, and delivery of content knowledge that affect student learning.

Summative Evaluations – used to make final (usually end-of-year) decisions regarding teacher performance.

Teacher Effectiveness – the degree of impact teachers have on student performance.

Teacher Efficacy – a belief in one's ability to bring about desired outcomes.

Value-Added Growth Models – statistical methodologies that measure subject-specific student growth based on student scores on standardized achievement tests.

Chapter 2

Literature Review

While the vast majority of educational literature cites the most critical factor for improving student achievement is the effectiveness of the classroom teacher (Danielson, 2008), there is significant variability and even controversy regarding how the effectiveness should be measured. Each of the proposed methods has some critical deficiencies to overcome. In light of the parallel movement to use these tools for high-stakes' decisions regarding teacher placement, promotion, and dismissal, ensuring the accuracy and reliability of these methods is paramount. A review of the literature tracing the evolution of teacher evaluation purposes and methodologies, the meaning of teacher effectiveness, and the strengths and weaknesses of various evaluation protocols is presented. In addition, recent and relevant findings regarding the three-year pilot of the largest research project ever conducted in teacher evaluation and its impact on a new state-wide teacher evaluation system at the center of the proposed research is investigated.

Accountability

As knowledge of the gaps in achievement among disadvantaged students in this country came to the forefront of public education issues, the No Child Left Behind (NCLB) legislation of 2001 was drafted with a focus on increasing the accountability of schools to set high standards and establish measurable goals to improve individual student achievement (Kachur, Stout, & Edwards, 2010; Pallas, 2012). As a result, NCLB required districts to staff classrooms with highly qualified teachers, based on research that indicates "nothing is more important to high achievement as having effective teachers" (Hanushek & Rivkin, 2010, p. 133). With the simultaneous advancements in technology enabling the collection of massive amounts of student assessment data, accountability for achievement is now being attributed directly to individual teachers. However, the ability to link student test scores to specific teachers does not necessarily validate the use of standardized test scores as the sole measure of teacher effectiveness. There are many factors affecting student scores (e.g., classroom size, availability of resources, mobility of students, to name a few) that cannot be accounted for in the achievement data, as well as other important (and immeasurable) contributions teachers make in the daily lives of their students (caring, compassion, connections with students, etc.).

Federal involvement in teacher evaluations. The serious political and economic problems attributed to the gap in student achievement across this country have led to a "rapid expansion of systems intended to hold schools and teachers accountable for student performance" (Pallas, 2012, p. 54). In response to the NCLB requirement that highly qualified teachers were to be placed in every classroom by the 2005-06 school year, the National Governors Association (NGA) identified six policy goals for improving student learning: "define teacher quality, focus evaluation policy on improving teacher practices, incorporate student learning into teacher evaluation, create professional accountability through career ladders, train evaluators, and broaden participation in evaluation designs" (Goldrick, 2002, p. 1). States adopting these policy goals express the belief that embedding these strategies into state and policy regulations will improve student achievement (Hazi & Rucinski, 2009).

Two additional federal initiatives, the Race to the Top (RTTT) and the Elementary and Secondary Education Act's (ESEA) Flexibility Programs, have "triggered a remarkable overhauling of the nation's teacher-evaluation programs" (Popham, 2013, p. 19). The $4.5 billion funding available in the 2009 RTTT program lured many states into raising their standards for teacher evaluation and including gains in student achievement as a significant factor in the evaluation in order to qualify for the funding (Alvarez & Anderson-Ketchmark, 2011; Popham, 2013; Schachter, 2012). In 2010, the first two states awarded RTTT grants included the Danielson framework in their proposals (Alvarez & Anderson-Ketchmark, 2011). In 2011, states were offered the ability to apply for the ESEA Flexibility program, which provided a federal waiver to the NCLB sanctions, but also required major changes to the state's system of teacher evaluation (Popham, 2013). Unsuccessful in its initial application for RTTT funding, Pennsylvania submitted an updated application during Phase 3, maintaining commitment to increasing teacher effectiveness. Pennsylvania received an award in excess of $41 million in April 2011 (United States Department of Education [ED], 2011). In June 2012, Pennsylvania's legislature passed Act 82, mandating an overhaul to the teacher evaluation system to mirror those implemented in other RTTT states.

Impact. The federal push for teacher evaluations is creating tension between teachers and teachers' unions (Schachter, 2012) and has raised concerns that a rush to implementation by states "could have serious adverse effects" (Phillips & Weingarten, 2013, p. 24), while providing no help for teachers to improve their performance (Mielke & Frontier, 2012, p. 10). Many school systems are using the revamped teacher evaluation tool "as a giant sorting mechanism whose purpose is to rank and rate teachers, bestow bonuses and other extrinsic benefits on the high flyers, and target the low scorers for remediation or dismissal" (Simon, 2012, p. 61). The potential use of the evaluation process to improve instructional practice is likely to have far more positive impact on student achievement than using the process primarily as a summative rating tool on teachers.

Teacher Effectiveness

Since the early twentieth century, questions regarding teacher effectiveness have been of primary importance to the field of education (Doyle, 1977). In 1966, the seminal Coleman Report indicated schools and teachers have a limited ability to improve student achievement (Rivkin, Hanushek, & Kain, 2005). However, the Coleman research did not take into consideration factors that can influence teacher effects, such as nonrandom assignment of students. Large-scale studies provided convincing evidence that teachers do indeed make a difference if ways to measure these factors are incorporated into the research (Haycock, 1998). Rivkin et al. (2005) used a fixed effects model to control "explicitly for student heterogeneity and the nonrandom matching of students, teachers, and schools" (p. 418), and found large differences in teacher quality within schools. A significant finding was that little variation in teacher quality was due to differences in their experience or levels of education (Rivkin et al., 2005).

Starting in 1960, research on teacher effectiveness relied on observational methods attempting to correlate student achievement with teacher behaviors (Westbury, 1988) and early researchers hypothesized that "certain teaching acts and conditions would affect student outcomes" (Seidel & Shavelson, 2007, p. 455). Known as the process-product paradigm, this framework investigated the relationship between teacher classroom behaviors and student learning outcomes (Westbury, 1988). Product variables are the student outcomes; process variables are the teaching approaches that lead to student outcomes (Seidel & Shavelson, 2007). Other major influences on student outcomes are context variables, including factors such as parental involvement, availability of technology, and student demographics (Seidel & Shavelson, 2007).

There is substantial research supporting the supposition that the academic qualities of the teacher (including content knowledge, verbal ability, and math skills) impact teacher effectiveness (Council for the Accreditation of Educator Preparation [CAEP], 2013). Likewise, non-academic qualities of the teacher (e.g., communication skills, perseverance, focus, ability to motivate) are thought to be associated with teacher effectiveness (CAEP, 2013). However, no empirical research exists that identifies or measures these non-academic qualities (CAEP, 2013).

The use of performance criteria to evaluate teacher effectiveness simplified the collection of measurable and observable teacher behaviors and was successful in "producing an accumulation of findings linking teacher behavior to student achievement" (Westbury, 1988, p. 147). Several aspects of the process-product paradigm have been criticized, specifically, the use of standardized testing as the only measure of student achievement, the correlational nature of the research (Grant & Drafall, 1991), and the lack of theoretical grounds, primarily due to "methodological problems that have impeded attempts to compare studies, integrate findings, or apply results to teacher education" (Doyle, 1977, p. 164).

An interesting finding was that teacher effectiveness is linked to the teacher's perceived sense of efficacy, which refers to a teacher's belief in his or her ability to affect student performance (Guskey, 1987; McLaughlin & Marsh, 1978). Three variables are thought to impact the teacher's sense of efficacy: (a) the student's performance, (b) the student's ability, and (c) the teacher's scope of influence (Guskey, 1987). The self-efficacy of teachers of high performing students was found to be greater than those with low performing students; teachers of students with low abilities reported a lower degree of self-efficacy than teachers of high-ability students (Guskey, 1987), and have been found to be less attentive to their low-ability students (Brophy & Evertson, 1977). The third variable, scope of influence, refers to the differences in self-efficacy teachers expressed for groups of students compared to individual students:

> When poor performance was involved, teachers expressed less personal responsibility and efficacy for single students than for results from a group or entire class of students…. Poor performance on the part of a single student was generally attributed to situational factors outside of the teachers' control. (Guskey, 1987, p. 46)

Although the primary tool for judging the effectiveness of teachers has been some model of teacher evaluation, the majority of evaluation systems fail to make a distinction among teacher performances in that more than 99% of teachers receive a satisfactory rating (Kane et al., 2011; Weisberg et al., 2009). In essence, teacher effectiveness is largely ignored in all but the most egregiously poor performances (Weisberg et al., 2009).

Teacher Efficacy

Teachers' self-efficacy beliefs about their competence in a given situation affect many important educational outcomes, including their investment in the teaching profession, levels of aspiration, persistence in the face of challenge, and resilience when experiencing setbacks (Tschannen-Moran & Hoy, 2001). Teachers' perceptions of their own abilities and the contexts in which they teach both influence and are influenced by their environment, which in turn is thought to affect their classroom behaviors (Fives & Buehl, 2010; Tschannen-Moran & Hoy, 2001). For example, some teachers believe that external factors hinder their ability to impact student learning, while other teachers "who express confidence in their ability to teach difficult or unmotivated students evidence a belief that reinforcement of teaching activities lies within [their] control" (Tschannen-Moran, 2001, p. 784).

Research on teacher self-efficacy has explored relationships between teachers' sense of self-efficacy and important outcomes related to teacher performance and student achievement (Heneman, Kimball, & Milanowski, 2006). A longitudinal analysis of teacher self-efficacy as a "predictor of subsequent teacher classroom performance and student value-added learning" (Heneman et al., 2006, p. 4) used the Danielson *Framework for Teaching* to measure teacher performance. Two important findings were reported: (1) the teacher self-efficacy scores measured at the beginning of the year were found to be significantly related to the end-of-year teacher performance ratings, and (2) these scores were not significantly related to ratings of student achievement (Heneman et al., 2006). It was surmised that any impact of teacher self-efficacy on student achievement "would be mediated by its impact on teacher performance" (Heneman et al., 2006, p. 12). Furthermore, Heneman et al. (2012) suggested a meta-analysis be performed on prior studies showing a positive link between efficacy and student achievement, in order to determine if control for teacher performance was included. To date, no further research on this relationship has been reported.

Teacher Evaluation

The focus of teacher evaluation has changed over the years, aligned to continuously evolving research on teaching and student learning. Until recently, evaluation systems have remained primarily a function of local initiatives. The purpose of teacher evaluations has also vacillated between a formative tool to improve instruction and a summative tool to provide end-of-year ratings of teachers.

History. In 1965, the ESEA introduced the concept of educational program evaluation, requiring districts receiving federal monies to evaluate mandated programs (Popham, 2013). A great deal of effort to change the perception of teacher supervision from an evaluative to a more collaborative function occurred during the 1960s and 1970s (Hazi & Rucinski, 2009).

In the 1970s, a movement toward open education was taking hold, whereby students were given a choice of tasks, able to move freely around the room, and receive individualized, self-paced instruction. At the same time, methodological advances in research design provided a scientific basis for linking teaching behaviors to student learning (Brophy, 1979). The key findings generated in these studies were (Brophy, 1979):

- Direct instruction is an effective method for producing student learning of basic skills;
- There is no support for the notion of generic teaching skills…Few, if any, specific teaching behaviors are appropriate in all contexts;
- Students taught with a structured curriculum do better than those taught with individualized or discovery learning approaches;

- Those that receive much instruction directly from the teacher do better than those expected to learn on their own or from one another; and
- Teacher talk in the form of lectures and demonstrations is important, as is the time-honored methods of recitation, drill, and practice (p. 18).

Other early research on the teaching behaviors related to academic success was reported by Rosenshine (as cited in Darling-Hammond et al., 1983), finding that direct instruction with frequent, single-answer questioning, large-group instruction, and controlled practice provide the best results in student achievement, while the "use of higher order, divergent, or open-ended questions, exploration of students' ideas, student-initiated discourse or choice of activities, conversation about personal experience or about subject matter tangential to the immediate objectives of the lesson at hand" (p. 295) should be avoided.

Within 20 years, and as a result of extensive research studies on effective classroom instruction, the entire perspective on effective instructional strategies has changed. Marzano, Pickering, and Pollock's (2001) report on research conducted by the Mid-continent Research for Education and Learning (McREL) to "identify those instructional strategies that have a high probability of enhancing student achievement for all students in all subject areas in all grade levels" (pp. 6-7) stood in direct contrast to the Brophy and Rosenshine findings. According to the McREL research, the categories of instructional strategies with the highest effect sizes on student achievement include

- Identifying similarities and differences;
- Summarizing and note taking;
- Reinforcing effort and providing recognition;

- Homework and practice;
- Nonlinguistic representations;
- Cooperative learning;
- Setting objectives and providing feedback;
- Generating and testing hypothesis; and
- Questions, cues, and advance organizers. (Marzano et al., 2001, p. 7)

McNeil and Popham (as cited in Darling-Hammond et al., 1983) argued that teacher evaluations should be based on their contributions to student performance, one of the first references to using student scores on standardized tests as valid measures of teacher effectiveness. Teachers, on the other hand, opined that student results are impacted by additional factors beyond the control of the teacher (e.g. innate ability, demographics, class size, etc.). This sentiment was upheld by the Beginning Teacher Evaluation Study (BTES), in which researchers found that connections between teaching behaviors and student learning were not possible at the time (Darling-Hammond et al., 1983).

The linkage of specific teaching behaviors to student learning generated a focus on a new model of teacher evaluation – clinical supervision (Hazi & Rucinski, 2009). Before the 1980s, teacher evaluation remained a strictly local endeavor (Veir & Dagley, 2002). The first stirrings of state-sponsored teacher evaluation initiatives occurred after the release of *A Nation at Risk* in 1983. These earliest models included specific criteria, procedures, and instruments for the evaluation of teacher performance and training of evaluators (Hazi & Rucinski, 2009). As a protection against unannounced visits from administrators into classrooms for evaluative purposes, the pre-conference emerged as a clinical supervision practice endorsed by teachers and their union sponsors (Hazi & Rucinski, 2009).

Despite the emergence of a national call to implement standards-based reform in education, including teacher participation in professional learning, setting high expectations for all students, and altering their instruction in significant ways, tenured teachers were mostly left alone during this period (Youngs, 2013). As a result, they had little reason to change their practices: "the status quo was acceptable and their annual evaluations were very likely to be the same from one year to the next regardless of any efforts they made to improve their teaching – certainly not a formula for driving change" (Youngs, 2013, p. 11). Teacher evaluation continued to consist of classroom observations by an evaluator, typically consisting of a checklist of teacher behaviors and classroom conditions, followed by a brief meeting with the teacher to review the results. These clinical supervision models were criticized by practitioners and policymakers for several reasons: (a) they lacked any variation in performance levels – the vast majority of teachers received satisfactory ratings, (b) the focus of the observation instruments were only on generic teaching behaviors, and (c) the ratings had virtually no impact on classroom instruction (Youngs, 2013, p. 12).

Within the No Child Left Behind Act (NCLB) is the mandate that states and districts who receive federal funds address the lack of highly-qualified teachers in schools with substantial numbers of disadvantaged students (Peske & Haycock, 2006). Although the legislation went into effect in 2002, the US Department of Education did not enforce compliance with this provision until July 2006 (Peske & Haycock, 2006). A body of research conducted in three states by the Joyce Foundation found evidence that differences in teacher quality have a significant impact on student achievement (Peske & Haycock, 2006). Relevant characteristics of highly-qualified teachers include the following areas:

- Academic skills and knowledge – the teacher's level of literacy accounts for the greatest amount of variance in student achievement;
- Mastery of content – teachers with a major in the subject, particularly math and science, typically produce higher performing students;
- Experience – student performance typically improves after the teacher's first few years in the field; and
- Pedagogical skill – research varies, but licensure has been weakly correlated with quality. (Peske & Haycock, 2006, pp. 8-9)

Recent changes in federal legislation (NCLB, RTT, ESEA Flexibility Program) have brought forth "new approaches to teacher evaluation [that] focus much more on instruction, subject matter, and/or teachers' effects on student learning than did past teacher evaluation practices" (Youngs, 2013, p. 2). However, the identification and measurement of effective teaching practices are major concerns for all stakeholders (Kane, Taylor, Tyler, & Wooten, 2011). Furthermore, despite these uncertainties, "policymakers want to treat the evaluation measures as though they are infallible and use them to place teachers in rigid boxes, labeling them as good teachers or poor teachers" (Pallas, 2012, p. 56).

Purposes of evaluation. There are two major purposes for teacher evaluations: (1) formative evaluations provide information that can be used to improve teacher effectiveness, and (2) summative evaluations provide information used to develop year-end evaluations and decisions concerning tenure and termination (Danielson, 2008; Popham, 2013). Some policymakers and state educational department advisors are promoting the need to

> use teacher evaluations for *selection* – to weed out ineffective teachers and perhaps identify the best ones for rewards, such as merit pay. Others view teacher evaluation as a tool for *direction*, pointing teachers toward aspects of their classroom practice that they can improve. (Pallas, 2012, p. 54)

Inherently different systems are needed for an evaluation system focused on developing teachers and improving learning than a system focused on measuring teacher competence (Marzano, 2012b), and difficulties arise "in integrating the requirements of an evaluation policy geared toward job status decisions with those of a policy aimed at improving teaching" (Darling-Hammond, Wise, & Pease, 1983, p. 287). Defining and evaluating teacher effectiveness must rely on valid instruments that can hold up to judicial scrutiny, and involve the development of "reliable, generalizable measures of teaching knowledge or behavior" (Darling-Hammond et al., 1983, p. 287). The more limited the evaluation criteria are for summative purposes, the less value it has for formative purposes (Darling-Hammond et al., 1983). Marshall (2005) gave important reasons why typical evaluation processes are not effective models to improve teaching. One observation is only a fraction of the total amount of a teacher's instructional year, and it is difficult for an evaluator to see all the nuances of instructional practices that occur even during that one lesson.

In addition, what occurs during that one lesson does not reflect the sum total of the impact a teacher may have on other important, often immeasurable, aspects of a child's education. Due to the advanced notice of the observation, observed lessons are often atypical and contrived performances. Results from evaluations that focus only on teaching performances may not reflect the impact of the instruction on student learning: just because something is taught does not mean it is learned.

As a major component of a teacher's summative evaluation, formal observations generate anxiety and make it difficult for the evaluator and teacher to openly discuss areas of improvement (Marshall, 2005). Revelation of instructional weaknesses to an evaluator who may incorporate the information into the teacher's summative evaluation is unlikely to foster honest, open communication (Popham, 2013). Teaching is a practice that can continue to improve throughout a teacher's career; however, "if the school views the need for improvement as a liability, why would teachers ever acknowledge their need for deliberate practice?" (Mielke & Frontier, 2012, p. 12). In most cases, an overall satisfactory rating is given, without providing teachers with specific information on how to improve. As a result, teachers rarely change their instructional practices.

Another concern regarding the dual purposes of evaluation revolves around the administrator, who must balance the role of an instructional coach and a supervisor. While coaches seek to improve the instructional practices of their teachers through honest discussions of areas in need of improvement, the administrator's supervisory role is charged with determining a final, summative evaluation for the teacher. While some districts may be able to use different personnel to serve in the coaching and supervisory roles, often the same administrators are mandated and challenged to simultaneously assume both roles (Hazi & Rucinski, 2009; Popham, 2013; Weber, 2012).

Fortunately, the preparation and professional development programs for principals in most states embrace the Interstate School Leaders Licensure Consortium (ISLLC) standards for instructional leadership. Each of the six standards describes the knowledge required for the standard, the dispositions manifested by accomplishment of the standard, and the observed performances expected of an administrator who has attained the standard. Although not explicitly mentioned in the ISLLC standards, enhancing the professional growth of teachers is a component of one of the standards. With the high-stakes attached to teacher evaluation, it is more important than ever for principals to be trained on the formative and summative aspects of these evaluations.

Evaluation Models

The changes in the purpose for teacher evaluations have been accompanied by shifts in the focus of evaluation models. Prior to the era of accountability, local school districts maintained control of both the focus and the manner of teacher evaluation, which mostly consisted of traditional observations of teachers. Notably, there were no high-stakes attached to the results. With the availability of detailed student achievement data and advances in statistical modeling, new modes of evaluation are being implemented and mandated in states in response to the public demand for higher teaching standards.

Traditional models of evaluation. Until July 2013, there was no mandate for tenured teachers in Pennsylvania to be rated with more than an annual evaluation resulting in a designation of satisfactory or unsatisfactory. The evaluation could, but was not required, to include a classroom observation conducted by an administrator. Marshall (2005) provided 10 reasons why the traditional/clinical supervision process consisting of one annual classroom observation is unlikely to improve teaching or learning:

1. Principals evaluate a tiny fraction of a teacher’s yearly classroom instruction. One period of instruction is approximately 0.1% of the instructional year (p. 728);
2. Evaluations based on one lesson do not incorporate the myriad of important components of a teacher’s professional responsibilities (lesson planning, curriculum development, grading, parental outreach, professional development, etc.);
3. Announced observations often result in atypical lessons, the “old dog and pony show” (Pieczura, 2012, p. 72), carefully designed to match the expectations of the evaluator;
4. Few instructional practices are likely to be observed in one isolated lesson;
5. Student learning, the most important outcome of instructional practices, is not represented in the traditional clinical observation;
6. Formal evaluations are high-stakes, often anxiety-producing experiences for the teacher, making it difficult for open dialogue about deficiencies to occur;
7. The clinical observation fails to address a “prevalent tendency in many schools: ‘I taught it, therefore they learned it’” (p. 730);
8. Evaluation instruments, designed to capture all the subtleties of classroom instruction, can hinder the principal’s ability to obtain a holistic view of the instruction. Good teaching is extremely complex and challenging (Lavy, 2007; Phillips & Weingarten, 2013; Scherer, 2012; Seidel & Shavelson, 2007). Even an experienced observer will have difficulty in analyzing and

> simultaneously documenting the most important elements of effective classroom instruction;
>
> 9. Most traditional evaluations require an overall rating (satisfactory/unsatisfactory), which inhibits in-depth review of the teacher's competency on specific performance standards and feedback on how performance can be improved; and
> 10. The hectic schedule and vast areas of responsibility assigned to principals limit the amount of time available for a comprehensive system of evaluation and supervision. Effective and efficient strategies for improving teaching and learning through the supervision process are not provided in traditional evaluation models. (pp. 728-731)

The traditional model of teacher evaluation assumes student learning will improve if we accurately rate teachers (Bambrick-Santoyo, 2012). There is no evidence that traditional, clinical observations of teachers have been able to produce these desired improvements in student achievement. Bambrick-Santoyo (2012) suggested "The core driver of teacher development is not accurate scoring, but skillful coaching, working with instructors on specific concrete actions that will improve results" (p. 27).

Using student achievement for evaluation. As a result of standards-based reform movements, common instructional standards have proliferated our K-12 instructional institutions. The initial challenge for school leaders was finding methods to determine if their programs were meeting these standards. Initial attempts to evaluate the effectiveness of instructional programs relied on the use of standardized tests to provide student achievement data. Unfortunately, "there are limitations and flaws inherent in relying heavily on test scores" (Pickering, 2012, p. 1). Notably, year-end achievement results are not available in time for teachers to adjust their instruction, they do not provide diagnostic information on individual students, and the scope of what they assess is limited (Gates, 2012).

With advancements in technology, student achievement data can now be linked to specific teachers and have become a focal point for measures of teacher effectiveness. This is due, in part, to the lack of "reliable and valid information on [teacher] effectiveness through direct observation of teachers in the act of teaching" (Kane et al., 2011, p. 56). However, there are significant areas of concern with the use of student test scores for this purpose.

By reducing the determination of teacher effectiveness to evaluating an educator's impact on student achievement as measured by standardized tests, other ways that teachers contribute to the success of their students is overlooked (Goe, Bell, & Little, 2008). "Similarly, other influences on student outcomes, including teachers in previous grades, specialists, peers, school resources, tutoring, community support, leadership, and school climate or culture cannot be 'parceled out' of the resulting score" (Goe et al., 2008, p. 5). If true teacher effectiveness is a culmination of all the processes and outcomes deemed important for students to experience, then "multiple measures – each designed to measure different aspects of teacher effectiveness – must be employed" (Goe et al., 2008, p. 51). The incorporation of various data sources should weight those factors that most accurately measure student progress and achievement (TNTP, 2010).

Value-added models. The latest trend in evaluation protocols is the inclusion of powerful statistical value-added models (VAMs) that can measure annual student growth in selected subjects and grade levels. At this point, 22 states have incorporated student growth in achievement as a significant component of teacher evaluations (Schachter, 2012). The use of value-added gains or losses for measuring teacher effectiveness must be validated as directly caused by the effectiveness of the teacher and not attributable to external factors beyond his control. Corcoran (2010) pointed out that the potential of using value-added assessments to create a school environment "in which teachers and principals work constructively with their test results to make positive instructional and organizational changes" (p. 7) to ultimately improve student achievement is extremely attractive. However, Corcoran cited significant concerns regarding the valid use of these models that need to be addressed: "What is being measured? Is the measurement tool appropriate? Can a teacher's unique effect be isolated? Who counts? Are value-added scores precise enough to be useful? Is value-added stable from year to year?" (p. 3).

Value-added proponents claim their methodology controls for the influence of variables that can impact student growth, such as socioeconomic status, disabilities, and levels of English proficiency (Caldas, 2012). While many educational scholars strongly support the use of VAMs to evaluate educators, others have significant concerns about using VAMs for summative teacher evaluations (Newton et al., 2010; Youngs, 2013). Notably, claims regarding the ability of VAMs to control for effects of multiple teachers (Lavy, 2007), student aptitudes, home environments, and family support are based on "untested statistical assumptions" (Newton et al., 2010, p. 1). The most controversial aspects of this methodology include concerns about the validity (Corcoran, 2010), reliability, and stability of value-added measurements, the non-randomness of student assignment to teachers, and various limitations inherent in these models (Newton et al., 2010; Youngs, 2013).

Validity and reliability of value-added measurements. The validity of basing the evaluation of teacher effectiveness on value-added models has been rejected by most members of the psychometric and educational research communities (Caldas, 2012; Newton et al., 2010). Since value-added predictions of individual student achievement are obtained from a model calculated with data collected from thousands, or even hundreds of thousands of individuals, there is typically a large margin of error making it invalid for accurate assessments of one student, one teacher, or even one school (Caldas, 2012, p. 2). An investigation on the validity of the Tennessee Value-Added Assessment System (TVAAS), now an integral component of many statewide teacher evaluation models, revealed concerns from stakeholders with the interpretations and attributions of the causes of student results obtained from the complex calculations (Kupermintz, 2003). Proponents of TVAAS methodology claim that the use of student prior achievement as a covariate “adequately accounts for all the potent external influences on student learning, thereby allowing the proper isolation of teacher direct effects on learning” (Kupermintz, 2003, p. 294). Kupermintz (2003) countered this claim, concerned that “failure to achieve proper isolation of teacher direct effects on learning may result in perverse policy decisions, benefiting teachers who are routinely assigned to students likely to make stronger gains” (p. 294), whereas teachers servicing at-risk student populations are subject to negative evaluations.

Non-randomness of student assignments. A significant concern about value-added measures is the potential for unfair evaluations of teachers as a result of the students they teach rather than their actual impact (Rothstein & Mathis, 2013). The assignment of students to teachers is rarely done on a random basis (Goodwin & Miller, 2012); instead, many observable (e.g., demographic data, prior achievement) and unobservable characteristics (e.g., influence of peers, motivation) of students can affect their achievement, which eliminates the possibility that any one teacher's class load is a random representation of students (Youngs, 2013). There is tremendous variability in the nature of teachers' assigned rosters. Some may have significant numbers of low-performing or high-performing students, perhaps by choice, chance, or seniority considerations. Despite claims that VAMs control for such instances, research reveals that teachers with larger numbers of disadvantaged students are more likely to be rated as less effective (Newton et al., 2010; Rothstein & Mathis, 2013). Unfortunately, the bias revealed by this research may have an undesirable impact: the reticence of teachers to embrace challenging class assignments that will negatively impact their formal evaluations (Bennett, 2012; Rothstein & Mathis, 2013).

Other limitations. Value-added measures can only be applied in grade levels and subjects in which state-wide, standardized assessments exist, approximately 25% of K-12 teachers (Kane et al., 2011). Furthermore, other than purporting to determine teacher effectiveness, such measures provide no information to guide professional development or highlight best practices responsible for successful teacher ratings (Kane et al., 2011). Like other measurements that rely on student assessment data, a focus on test-taking skills, test preparation, narrowing of curricular content, and elimination of equally valuable, but untested content, is likely to result (Kane et al., 2011).

Recommendations. While the traditional/clinical observation evaluated the inputs (lesson plan, classroom preparation, teaching methods, classroom assignments, questioning techniques, etc.) provided by teachers during the course of the year, the statistical models evaluate the outputs (growth in achievement of students as measured by standardized test scores). Some may view traditional observations as obsolete, biased, and ineffective for making summative decisions regarding teacher effectiveness, yet the value-added growth models do not provide educational leaders with information that will reveal the strengths and weaknesses of classroom instruction. Just as critical, value-added measures cannot help educators understand why certain practices or behaviors are more successful than others. Although value-added scores can be used by educators to provide diagnostic information on individual or cohorts of students, they do not provide evaluators or teachers with information on improving specific classroom practices (MET, 2010). In non-test subjects and grade levels, value-added models are not available on which to base any decisions regarding teacher effectiveness. As a result, teachers performing at moderate levels will not get the critical professional development and differentiated support necessary to improve their effectiveness (Weisberg et al., 2009).

Value-added measures provide important information for the assessment of a school's programs and the progress of its students, taken in aggregate, towards meeting state-wide measures of proficiency. Caution is warranted, however, in using these scores for the purpose of summative measures of teacher effectiveness. Value-added scores may "reasonably be considered as one component of teacher evaluation" (Youngs, 2013, p. 21), but they should be combined with other sources of data that are considered relevant to the school's beliefs about effective teaching (Goodwin & Miller, 2012). DiCarlo

(2012) cautioned: “there is virtually no empirical evidence as to whether using value-added or other growth models … in high-stakes evaluations can improve teacher performance or student outcomes…it has never really been tried before” (p. 38).

With the potential high-stakes impact of student test scores on individual teachers, unintended and undesirable classroom practices have emerged. Teachers may opt to focus strictly on the tested standards, skimming over or eliminating important content that is not being evaluated on the test. Test prep frenzy has appeared in many schools, often for months prior to the state assessments (Marshall, 2012). Furthermore, a disturbing trend has emerged in recent years: accounts of cheating by individual teachers, administrators, and entire school systems have been reported (Lavy, 2007).

While the accountability mechanism of NCLB, Adequate Yearly Progress (AYP), supposedly uses test scores to objectively evaluate schools, the disaggregation of assessment data can result in unfair evaluations of the school’s true effort to reduce gaps in student achievement (Martin, 2011). This is a significant issue in schools with large populations of English Language Learners (ELLs) and students with disabilities.

By definition, ELL students lack proficiency in the English language. However, their assessment scores in content-specific areas “will in effect function as a test of English proficiency…[which] is not what is supposed to be measured when evaluating whether a school is making adequate yearly progress” (Martin, 2011, p. 11). Only districts with significant numbers of these students will be responsible for this subgroup’s achievement, unfairly penalizing the school for serving this population.

NCLB mandates the inclusion of students with disabilities in the assessment process to ensure accountability for their achievement. Schools with significant numbers of students with disabilities will be unfairly punished via AYP designations for serving this population. Improving the academic achievement of these students is critical to their futures, but

> given the heterogeneity of types and degrees of disability and given the nature of many prevalent processing disorders, there is evidence of a wide range inconsistency, and unpredictability of group scores that makes the validity of evaluating schools with students with disabilities using AYP measures highly questionable. (Martin, 2011, p. 12)

Prior research has determined that the effects of highly effective (or highly ineffective) teachers on the academic achievement of students persist for several years (Sanders & Horn, 1998). As a result, mediocre teachers may benefit from the halo effect of their students' exposure to highly effective teachers, and conversely, the value-added scores of highly competent teachers may be negatively impacted by their students' prior presence in the classrooms of ineffective teachers (Sanders & Horn, 1986; Goodwin & Miller, 2012).

Standards-Based Evaluation Models

The use of standards-based teacher evaluation models to measure teacher performance has gained tremendous momentum over the last 15 years. Recent federal legislation is experimenting with teacher compensation systems based in part on new evaluation models of classroom instruction and focused on the premise that "school leaders can identify more effective teachers through performance evaluations" (Kimball & Milanowski, 2009, p. 35). In order to evaluate the performance of teachers, a comprehensive and robust evaluation system is necessary that "fairly, accurately and credibly differentiates teachers based on their effectiveness in promoting student achievement" (Weisberg et al., 2009, p. 5). Strong implementations have the potential to provide feedback likely to enhance teacher reflection and growth through the significant amount of discourse between evaluator and teacher (Kimball, 2002). However, there are many reported cases of weak implementation practices that result in a number of concerns regarding the sole use of this tool for measuring teacher effectiveness.

Observation protocols. Although classroom observations have been an essential component of teacher evaluation for many years, early protocols were not grounded in evidence-based models of effective teaching (Youngs, 2013). Newer observation models use rubrics that allow evaluators to differentiate the effectiveness of teachers across various levels of performance standards by producing a detailed account of what they observe during a lesson (Danielson, 2008; Danielson, 2012; Youngs, 2013). Effective evaluation protocols should clearly define what is considered good teaching and what will serve as evidence for each element of the model (Danielson, 2012). Observation protocols should use evidence that enables teachers and observers to identify teachers' strengths and areas for improvement, and provides specific information to school leaders regarding needs for professional development (Youngs, 2013). The two major standards-based evaluation models being implemented across the nation are: (1) Robert Marzano's *Teacher Evaluation Model*, consisting of 41 key strategies, and (2) Charlotte Danielson's *Framework for Teaching*, which encompasses 76 criteria to judge teacher effectiveness (Bambrick-Santoyo, 2012). Pennsylvania has selected Danielson's framework for its mandated teacher evaluation model. The four domains and respective components of the framework are

1. Planning and Preparation
   A. Knowledge of content and pedagogy
   B. Demonstrating knowledge of students
   C. Setting instructional outcomes
   D. Demonstrating knowledge of resources
   E. Designing coherent instruction
   F. Designing student assessments
2. Classroom Environment
   A. Creating an environment of respect and rapport
   B. Establishing a culture for learning
   C. Managing classroom procedures
   D. Managing student behavior
   E. Organizing physical space
3. Instruction
   A. Communicating with students
   B. Questioning and discussion techniques
   C. Engaging students in learning
   D. Using assessment in instruction
   E. Demonstrating flexibility and responsiveness
4. Professional Responsibilities
   A. Reflecting on teaching
   B. Maintaining accurate records
   C. Communicating with families

D. Participating in a professional community

E. Growing and developing professionally

F. Showing professionalism. (Danielson, 2011)

Empirical studies of models. Studies conducted by the MET Project, the Consortium on Chicago School Research (CCSR), and Cincinnati schools show that teachers' observation ratings on each of the components of the FFT is associated with gains in student achievement (Danielson, 2012; Youngs, 2013). The teachers' classroom observation ratings in the Cincinnati School System predicted the achievement gains of their students in both math and reading (Kane et al., 2011). Heneman, Milanowski, Kimball, and Odden (2006) reviewed large scale studies in which standards-based teacher evaluation models were implemented across four different school districts. Additional research was conducted with a focus on the designs and effectiveness of these systems in order to determine their potential to be acceptable, valid, and useful for basing knowledge- and skill-based pay initiatives. A strong positive relationship between teacher evaluation scores and student achievement were found; the results were stronger in districts that used multiple evaluators with extensive training. Interviews of teachers revealed positive reactions with regards to their understanding of the evaluation's standards and rubrics, agreement that the system's domains reflect good teaching, positive impact on some areas of instructional practice, such as lesson planning, reflection, and classroom management, and increased focus on student standards. Both teachers and administrators commented that professional dialogue was improved under the new system. Issues of concern included additional workload for both groups, implementation glitches that "provoked doubts about the validity and fairness of the system" (Heneman et al., 2006, p. 7), the need for extensive training of teachers and evaluators, lack of emphasis by evaluators on providing timely and useful feedback, and lack of alignment between the teacher evaluation and other district functions (e.g., professional development, induction, and administrative evaluation which holds administrators accountable for full participation in the process).

A specific finding of concern was the variation in the strength of the relationship between student achievement and teacher evaluation ratings using a new standards-based teacher evaluation system. Evidence of a positive relationship between the proficiency ratings of teachers and the gains in achievement levels of their students on state criterion-referenced tests was reported (Holtzapple, 2003). However, the evaluation system was found to be most sensitive at the extremes (able to identify unsatisfactory and distinguished teachers), but appeared to be less sensitive in discriminating between teachers rated basic or proficient (Holtzapple, 2003). Perhaps teachers exhibiting "proficient" skills during the observation are not using these skills on a day-to-day basis. Holtzapple (2003) found that some teachers who score proficient during an observation have less than expected student gains. School leaders often express their concern of an observation being just a "dog and pony show." Further research may explore additional criterion necessary to validate the designation of "proficient" in a system of teacher evaluation.

Other concerns regarding the relationship between standards-based evaluation scores and student achievement were reported in the literature. Milanowski (2004) determined that the evaluation scores of a teacher in one year did not accurately predict student achievement in subsequent years. A plausible explanation suggests that evaluation scores are not accurate reflections of teachers' performances, but are more consistent with the proposition that teachers put forth more effort in the year(s) they are comprehensively evaluated (Milanowski, 2004).

Without consistent correlations, unfair and invalid ratings may occur with individual teachers. Kimball and Milanowski (2009) studied differences in evaluator decision-making to determine plausible explanations for differential validity across principals. Three influences were investigated: evaluator motivation, evaluator expertise, and evaluation context (Kimball & Milanowski, 2009). Motivation may have an impact on the degree of evaluator leniency in that evaluators intent on maintaining good working relationships with teachers may be hesitant to provide negative feedback. The evaluator's own skills and content-specific knowledge may impact accurate ratings, particularly in areas different than the evaluator's training and certification. The context of the evaluation may impact the evaluator's perspective. For example, in schools with a large number of low performing teachers, inflated ratings, chiefly due to evaluator bias, are likely to occur, and vice versa in schools with high performing teachers (Kimball & Milanowski, 2009).

No individual factor (motivation, skill, or context) was found to explain the variation in valid ratings. Some teachers expressed concern about the nature of their feedback, which provided little constructive criticism or recommendations on specific teaching strategies (Kimball & Milanowski, 2009). Principals were cautious in writing evaluation summaries, concerned that negative criticism would become part of the teacher's permanent record; they reserved more formal and intensive evaluation summaries for the weakest teachers (Kimball & Milanowski, 2009). Overall, the research findings suggest caution in using principal evaluations for high-stakes decisions. In order to promote greater accuracy and consistency among evaluators, greater attention to training, oversight, and evaluation validity needs to be in place (Kimball & Milanowski, 2009).

Reliability of models. The accuracy and reliability of standards-based teacher evaluation models are concerns, particularly for teachers who may face nonrenewal as a result of poor ratings. For example, the evaluation system in one study was most sensitive at the extremes (able to identify unsatisfactory and distinguished teachers), but was less sensitive in discriminating between teachers rated basic or proficient (Holtzapple, 2003). Milanowski and Heneman (2001) determined that the evaluation scores of a teacher in one year did not accurately predict student achievement in subsequent years, while Holtzapple (2003) found that some teachers who score proficient during an observation have less than expected student gains.

The observation protocol is reliable if different observers “can agree on the level of quality of what they observe” (Danielson, 2012, p. 34). Since a single observation score is dependent on various classroom factors that may not be indicative of the teacher’s actual effectiveness, researchers suggest scores over multiple observations will improve the reliability of the evaluation (MET, 2012). Districts can avoid measurement error by ensuring evaluators are extensively trained on using the system (Marzano, 2012a; Pallas, 2012; Youngs, 2013), and how to differentiate between evidence, opinion, interpretation, and bias (Danielson, 2012; Schachter, 2012). Pennsylvania provided school administrators the opportunity to train on the Danielson framework through an online professional development platform known as Teachscape©. The program offers 30 hours of training, followed by a six-hour test to measure the administrator's proficiency in accurately rating teachers in each of the four domains.

Sampling error occurs when the observation does not represent a typical lesson for the teacher (Marzano, 2012a). Researchers have found that one observation annually does not provide a complete picture of a teacher's competencies (Marshall, 2012). A more "accurate representation of a teacher's pedagogical skill" (Marzano, 2012a, p. 82) can be obtained by increasing the number of times the teacher is observed (Pickering, 2012). The standard of reliability should be high when there are high-stakes' decisions attached to the observation; researchers in the MET (2012) project suggest that multiple lessons should be observed and scores averaged in order to reduce the "influence of an atypical lesson" (p. 2).

Use of surveys within protocols. Oversight of the implementation of a standards-based teacher evaluation model should consider the inclusion of teacher surveys, which can enhance and complement the process by efficiently capturing stakeholder feedback, improving teacher engagement and participation in the evaluation, providing an opportunity to foster teacher growth and professional development, uncovering obstacles in the implementation of the evaluation process, and promoting a school culture that fosters continuous growth (Simon, 2012; TNTP, 2010; Wiener & Lundy, 2013, p. iii).

Teacher perceptions regarding the fairness and consistency of the evaluation process are important indicators for district leader consideration. Inaccurate observations foster mistrust; ensuring accuracy requires intensive training of observers and measures of inter-rater reliability (MET, 2012). Asking for teacher input "creates or reinforces an expectation that these aspects will be addressed and provides the basis for assessing quality of implementation" (Wiener & Lundy, 2013, p. 5). Additionally, teachers are motivated towards the pursuit of excellence when they are able to objectively analyze their own practice and are provided an understanding of how to improve (Mielke & Frontier, 2012). Educational leaders can use the information obtained from surveys to improve the process, particularly during initial stages of implementation. According to Wiener and Lundy (2013), there are four potential topics to explore through teacher surveys: (1) fidelity of implementation, (2) impact of evaluation on teachers, (3) teachers' experience of support and development, and (4) teachers' overall impression of the evaluation system (p. 7).

Fidelity of implementation. Although the new standards-based evaluation system was considered a significant improvement over prior evaluation systems, Kimball (2002) suggested that the mere adoption of such systems does not guarantee transformation of the evaluation process without sufficient training and "accountability for evaluation quality and consistency" (p. 264). Teachers reported the new evaluation system used standards that were clearer than traditional models and the process allowed for their input into the evaluation process (Kimball, 2002). Discussions that focus on performance standards will enable teacher and observer to engage in more structured and consistent conversations (Goe, 2013).

Impact of standards-based evaluation on teachers. Teachers identified five areas of improved practice as a result of their participation in one standards-based evaluation model: (1) planning, (2) classroom management, (3) using assessment during instruction, (4) differentiated instruction, and (5) student-focused learning (Sartain, Stoelinga, & Brown, 2011, p. 27). Research on the impact of feedback on teachers' performances revealed all participants reported a substantial increase in the amount of discourse surrounding teacher performance; however, while some teachers reported a change in teaching practice as a result of the feedback, the changes were mostly related to "classroom interactions or organization, rather than to content-related instructional practices" (Kimball, 2002, p. 259).

Teachers' experience of support and development. One of the major concerns with any teacher evaluation system is the lack of quality feedback (Mielke & Frontier, 2012). Fostering open feedback between teacher and observer can lead to more effective teaching and provide an opportunity for them to reach their full potential (TNTP, 2012; Wiener & Lundy, 2013) and should be a goal of any supervisory model (Marshall, 2005).

Teachers' overall impression of the evaluation system. With regards to fairness, teachers' concerns focused on lack of consistency across different evaluators (Kimball, 2002). Perceptions of fairness include procedural fairness, interpersonal fairness, and outcome fairness. All three are believed to influence teachers' satisfaction and acceptance of the process and resulting evaluation appraisal, their motivation to improve, and trust in evaluators (Kimball, 2002). Kimball studied the experiences of evaluators and teachers in three school systems implementing a standards-based teacher evaluation model. Two of the focal points of the research were teachers' perceptions of the fairness of the process and teachers' assessment of the evaluation feedback (Kimball, 2002). Kimball (2002) suggests one method to improve feedback is to match teachers with evaluators who have similar content backgrounds. While feasible in larger school systems with a multitude of evaluators, this is impossible in districts with limited numbers of administrative staff.

Teachers' reactions to a pilot implementation of a new standards-based teacher evaluation system were captured to determine what objections, if any, should be considered and addressed prior to full implementation of the system (Milanowski & Heneman, 2001). Using interviews and survey methods, a correlation between teachers' overall satisfaction toward the system and various elements of the evaluation system were sought. While the teaching domains and standards were generally well understood and accepted by participants, the study's results revealed three serious challenges: (a) the process added too much to teachers' workloads, (b) sporadic and sparse feedback was given by evaluators, much of which was summative instead of formative, and (c) the system posed a threat to self-esteem or reputation (Milanowski & Heneman, 2001). To mitigate these concerns, evaluators need adequate time and training to appropriately use the system, including professional development on effective feedback techniques (Milanowski & Heneman, 2001). Teachers need to receive extensive orientation regarding the rationale and use of the system, and if teacher acceptance is deemed critical to the implementation, district leaders must carefully plan and attend to all facets of the process (Milanowski & Heneman, 2001).

Recommendations for use of protocols. The MET Project noted that observers face a tremendous burden when trying to track teacher competencies in the many components of standards-based evaluation models (Simon, 2012; Weber, 2012). One suggestion is to use multiple observers to capture more of elements considered to be essential to effective instruction and/or increase the number of observations on each teacher (Phillips & Weingarten, 2013). This will also help reduce the tendency of observers to base future ratings on their perceptions from initial evaluations (Sawchuk, 2013).

Announced classroom observations can prevent the observer from viewing the day-to-day classroom experiences of students. As a result, evaluations are not honest reflections of classroom interactions, “and are not helpful for improving mediocre and ineffective teaching practices” (Marshall, 2012, p. 50). The obvious solution is to schedule multiple unannounced visits to capture the most accurate representation of teacher effectiveness.

The system should delineate clear performance standards, include multiple rating options, ensure inter-rater reliability, and provide meaningful teacher feedback (Phillips & Weingarten, 2013; Weisberg et al., 2009). All evaluators should be thoroughly trained in the use of the evaluation model. Strong accountability standards applied to evaluators are essential, particularly in high-stakes decisions for teacher tenure, removal, and dismissal decisions. A major part of a successful evaluation model depends on the skill set, motivation, beliefs, and behaviors of the evaluator. Tomlinson’s (2012) description of the ideal evaluator captures most of these traits:

> works from the belief that no teacher ever finishes growing and everyone has the capacity to improve, frames feedback as a compliment to my capacity to grow in professional practice, calls attention to my strengths and helps me build on them, and delivers formative feedback and support *before* any summative evaluation. (pp. 88-89)

Any evaluation matrix should allow for distinct levels of teacher performance (TNTP, 2010). Evaluation systems should not be limited to a single rating; instead, frequent observations over time provide multiple opportunities for evaluators to gather a robust picture of teacher effectiveness. Evaluation processes that provide high ratings to all teachers "ignore(s) wide variations in effectiveness and stunt teachers' growth by giving them a false picture of their performance" (TNTP, 2012, p. 2).

The responses and behaviors of students reflecting varying levels of student engagement are rarely a focus in traditional teacher evaluations. Instead, evaluators typically use a check-list inventory of teacher-specific behaviors to determine teacher ratings. However, lack of student engagement is a chief indicator of teacher ineffectiveness, and should be a large component of any evaluation system (TNTP, 2010).

A high level of support should be available for teachers to meet these higher expectations for classroom practices (Phillips & Weingarten, 2013). The need for improvement should not be seen as a liability, but viewed as a component for teacher growth; otherwise, teachers are unlikely to openly discuss their need for support (Scherer, 2012). Admittedly, evaluation protocols also serve the purpose of weeding out teachers unable to improve their performance (Seo, 2012), but administrators must balance the two purposes of evaluation to best serve the needs of students.

Despite the legislative impetus for systems that evaluate teacher effectiveness, educational community experts cite the major premise behind standards-based evaluation models is their potential to improve student learning (Phillips & Weingarten, 2013). Charlotte Danielson says it best: "If we only get better at labeling teachers as underperforming or performing satisfactorily, we won't have done much for students" (Schachter, 2012, p. 44).

Measures of Effective Teaching (MET) Project

In the fall of 2009, the Bill & Melinda Gates Foundation launched the Measures of Effective Teaching (MET) project to gather information about the validity and reliability of several instruments used to measure teaching effectiveness, and to "identify effective teaching using multiple measures of performance" (Phillips & Weingarten, 2013, p. 24). Notably, the $45 million project (Sawchuk, 2013) was a massive random-assignment experiment intended to "eliminate systematic patterns in student assignments that might bias value-added scores" (Rothstein & Mathis, 2013, p. 2). Intent on exploring novel ways to recognize effective instruction, over 3,000 teachers in six urban districts across the country were involved in the two year research project (MET, 2010). The Gates Foundation spent 40 times the amount spent by the U.S. Department of Education on teacher effectiveness research (Gabriel & Allington, 2012).

Data sources included student achievement gains, classroom observations and teacher reflections, assessments of teachers' pedagogical knowledge, student input regarding classroom environment, and teachers' perceptions of instructional support. The primary difference between the MET project and research of teaching effectiveness over the last 50 years is the use of student test scores to judge effectiveness (Gabriel & Allington, 2012).

Classroom observation tools were found to provide accurate ratings if two or more lessons, scored by two different certified observers, were conducted on teachers (MET, 2013). Well-designed student surveys provided reliable feedback on teaching practices that was "predictive of student learning" (MET, 2013, p. 3). In January, 2012, the MET project researchers reported results on their two-year investigation of various classroom observation instruments, student survey tools, assessments of teachers' pedagogical knowledge, and different measures of student achievement (MET, 2012). While all observation tools were positively associated with gains in student achievement, several factors regarding the accuracy and reliability of the various instruments were noted (MET, 2012):

- One key finding concerns the accuracy of the ratings assigned by evaluators. A single observation score is dependent on various classroom factors that may not be indicative of the teacher's actual effectiveness. While a teacher's rating varied considerably from lesson to lesson, there was "little difference if those lessons were with different course sections" (p. 18) or with different groups of students. Hence, single observations are likely to produce inaccurate indicators of a teacher's classroom practice; instead, averaging scores over multiple observations will improve the reliability of the evaluation.
- Another obstacle uncovered in the research concerned the inter-rater reliability of the observers. One recommendation to address this concern is the implementation of training and certification of observers. Reliable evaluations of a teacher's practice should include multiple observations in order to capture an accurate picture of the large number of competencies and skills required of effective teachers. Reliability should be monitored by incorporating periodic observations by impartial observers.
- Combining observation scores with other factors improved predictability and reliability. For example, combining student feedback and student achievement gains was better than graduate degrees or years of teaching experience at

> predicting a teacher's student achievement gains with another group of students on the state tests and in identifying teachers whose students performed well on other measures. It is also important to periodically verify that there is a strong correlation between high observation scores and high gains in student achievement.

Quality implementation of a teacher evaluation project is a critical undertaking, with potentially significant impact on the educators and the system they serve. Six minimum requirements are essential: (1) an observation tool with clear expectations, (2) trained observers, (3) multiple observations, (4) double scoring of some teachers using an impartial observer, (5) triangulation of teacher ratings with observations, achievement data, and student input, and (6) periodic verification that high teacher ratings are aligned with high student achievement (MET, 2012).

Critical analyses of MET findings. Several researchers have reviewed the published MET research and offered some significant criticisms of the methodology and findings:

- The randomization protocol planned for the MET study was compromised as a large number of students either did not remain with their assigned teachers or did not complete the experiment (Rothstein & Mathis, 2013; Sawchuk, 2013);
- Many classrooms were not typical representations of classrooms across the district, in that students included in the randomization had higher baseline scores in math and reading than students not included (Rothstein & Mathis, 2013);

- Teachers with higher numbers of special education and English language learner students were not included in the study (Rothstein & Mathis, 2013);
- While the Gates Foundation considers value-added protocols as “the gold standard against which all else must be evaluated” (Gabriel & Allington, 2012, p. 44; Rothstein & Mathis, 2013, p. 4), qualitative research “that considers teaching as a multidimensional enterprise that serves a variety of purposes beyond test-score improvement” (Rothstein & Mathis, 2013, p. 5) was not addressed;
- The data did not sustain the project’s premise that the combination of value-added scores, classroom observations, and student surveys reflect one general teaching factor, nor should they be given equal weight in teacher evaluations. Instead, the review suggests each measure “captures a distinct component of teaching” (Rothstein & Mathis, 2013, Summary);
- Although the three measures predicted value-added student performance on multiple-choice-based state assessments, they did not predict student performance on assessments designed to capture higher-order thinking skills (Rothstein & Mathis, 2013). If these skills are considered valuable to a student’s education, “an evaluation system based around the MET measures will fail to identify teachers who are effective or ineffective on those other dimensions (Rothstein & Mathis, 2013, p. 9). Teachers and school systems focused on improving student achievement on the state assessments are less likely to focus instruction on more advanced, conceptual thinking skills (Rothstein & Mathis, 2013); and

- Although MET researchers claim that value-added scores are unbiased predictors of teacher effectiveness, the results do not adequately describe what constitutes effective teaching and offers “little guidance about how to design real-world teacher evaluation systems” (Rothstein & Mathis, 2013, Summary).

Pennsylvania Teacher Evaluation

Administrators in selected school districts across Pennsylvania completed three years of piloting the Danielson’s *Framework for Teaching* model as part of the governor’s proposal to mandate state-wide implementation. With high-stakes decisions likely to follow, evidence of the criterion-related validity of the PA model as well as evaluator preparation and inter-rater reliability are important considerations.

Effective with the 2013-14 school year, Pennsylvania formally adopted the Danielson *Framework for Teaching* evaluation protocol for all classroom teachers. The only districts not under this mandate are those who have existing contracts with their professional employees that designate other forms of evaluation. These districts must switch to the Danielson framework when their bargaining unit’s contract expires or renews. The protocol details five steps to the process:

1. The administrator provides the teacher with questions regarding planning and preparation for the targeted lesson (Domain 1 of the Danielson framework). The teacher and administrator then meet to discuss the teacher’s responses (Pre-Observation Conference).

2. The administrator observes the lesson, focusing on the 10 components within Domains 2 and 3 of the Danielson framework, and provides the teacher with a copy of collected evidence.
3. The administrator and teacher independently rate the evidence against the evaluative criteria for each component.
4. The administrator provides the teacher with a second questionnaire, which focuses on Domain 4 (Professionalism).
5. The administrator and teacher meet to compare their independent ratings for Domains 2 and 3. Together, they discuss the teacher's strengths and areas for growth.

The administrator assesses evidence of teacher proficiency according to a detailed, four-level scoring rubric points (Danielson, 2011; Kimball, 2002):

1. An unsatisfactory rating indicates the teacher's performance regarding the specific element does not meet minimum expectations and is given 0 points on the individual component;
2. A basic rating indicates basic understanding of the expected performance, but inconsistent or unsuccessful implementation, and is given 1 point;
3. A rating of proficient signifies both clear understanding and good implementation and earns 2 points; and
4. Distinguished ratings denote highly accomplished and mastery performance and provide the participant with 3 points.

While the legislation permits two models of evaluation (Formal Observation or Differentiated Supervision), both aligned to the Danielson framework, regular classroom walkthroughs are suggested to support administrators' summative ratings. Unannounced walkthroughs are recommended after the Formal Observation Model to verify the teacher's implementation of suggested improvements in classroom practices that arose during the post-observation conference. Walkthroughs are also recommended for all teachers in either evaluation model throughout the school year for both formative and summative purposes.

The alternative evaluation model in PA's legislation, Differentiated Supervision, may incorporate various suggested modes. These include, but are not limited to (PDE, 2013):

1. Peer Coaching Mode - professional employees work in dyads or triads to discuss and observe their own or another professional employee's pedagogy, student learning, curriculum aligned to the Pennsylvania Core Standards and other pertinent issues in a collaborative manner. The professionals will work together to define their professional needs and develop plans to assist them in the successful completion of the identified tasks including: specific target area(s), the evidence to be collected, observation dates, and a reflective session. Meeting notes, data collection tools, results of the observations, and the reflective sessions should be shared with the principal and used as evidence in the supervision and evaluation of the employee.
2. Self-Directed Model/Action Research Mode - professional employees will develop a structured, on-going reflection of a practice-related issue (Danielson *Framework for Teaching* or a PDE-approved alternative system). Professionals may work individually or in small groups, dyads or triads, to

> complete the action research project. Meeting notes, resources, data collection tools, and the results of the reflective sessions should be shared with the principal and used as evidence in the supervision and evaluation of the employee.
>
> 3. Portfolio Mode - professional employees will examine their own practice in relation to the Danielson *Framework for Teaching* or a PDE-approved alternative system and reflect in a written report and/or documented discussions with colleagues. Portfolios may be developed according to criteria established collaboratively by the administrator and the teacher based upon their interests or needs. Resources, data collection tools, and the results of the reflective sessions should be shared with the principal and used as evidence in the supervision and evaluation of the employee. (pp. 2-3)

Classroom Walkthroughs

One of the most important tools used by principals to gather information on what is occurring in their classrooms is the walkthrough observation (Kachur, Stout, & Edwards, 2010). Considered a standard practice for many years, walkthrough visits are used in a variety of settings to meet specific educational goals, and the selection of one protocol over another should be based on the needs of the particular school (Kachur, Stout, & Edwards, 2010). Generally, walkthroughs share a common characteristic: brief, informal visits conducted by administrators or other instructional leaders to capture snapshots of teaching and/or student performances over time (Kachur, Stout, & Edwards, 2010). Early classroom walkthroughs were implemented for supervisory and evaluative purposes (Fink & Resnick, 2001).In addition, classroom walkthroughs are mostly formative by nature, not intended for the formal evaluation of teachers (Kachur et al., 2010). When conducted on a regular basis, classroom walkthrough data can be used to illuminate how teachers make curricular and instructional decisions (Downey, Steffy, English, Frase, & Poston, 2004), ultimately help teachers analyze their teaching practices (Association for Supervision and Curriculum Development [ASCD], 2007), and improve student achievement (Kachur et al., 2010).

Purposes for classroom walkthroughs. Principals and other educational leaders use classroom walkthroughs for a variety of instructional and managerial goals by direct observation of actual practices. Effective principals are visible and accessible to teachers, students, and other school personnel, and walkthroughs of halls and classrooms provide opportunities for school leaders to observe and interact directly with these groups (Cotton, 2003; Kachur et al., 2010; Marzano, Waters, & McNulty, 2005). Walkthroughs help keep communication lines open and help school leaders monitor the impact of school practices on instruction (Marzano et al., 2005; Stronge, Richard, & Catano, 2008).

The principal's role as an instructional leader is enhanced by observing firsthand what is taking place throughout the school with regards to instruction and curriculum (Cotton, 2003; Downey et al., 2004; Kachur et al., 2010; Marzano et al., 2005; Whitaker & Zoul, 2008). Walkthroughs provide principals the opportunity to develop professional learning communities, working collaboratively with staff to reflect and analyze their own instructional practices (Cotton, 2003; Downey et al., 2004; Kachur et al., 2010; Stronge et al., 2008). With the national shift to common core standards, leaders are faced with evaluating the school's readiness and compliance with newly mandated curriculum (Kachur et al., 2010).

Walkthroughs provide evidence in the form of data for a number of important instructional purposes. Questions about student performance and teaching practices, identification of professional development needs of individual staff members, and progress of professional development initiatives are informed by data collected during classroom walkthroughs (Kachur et al., 2010; Stronge et al., 2008). Performance data on individual teachers supports the mentoring relationship with the evaluator, keeping him informed of the person's strengths and areas in need of improvement as well as uncovering obstacles impacting the teacher's performance (Cotton, 2003; Kachur et al., 2010).

The shift from a teacher-focused to a learner-focused supervision model requires principals to determine whether students are motivated and engaged during classroom instruction (Mandell, 2006; Kachur et al., 2010). Classroom visits are opportunities for the principal to determine whether indicators of student involvement exist (Gray and Streshly, 2008). Using multiple classroom visits focusing on learner behaviors facilitates and promotes discussion with teachers on classroom practices that contribute or detract from these desired student behaviors (Downey et al., 2004).

Practitioner views of walkthroughs. Principals report that teachers are more favorable of the evaluation process and hold a higher value of professional development after their participation in walkthroughs (Downey et al., 2004). Principals trained in various classroom walkthrough models expressed belief that the practice improved instruction and learning (Dexter, 2005). In another study of a new walkthrough observation tool, Keruskin (2005) reported that principals found teachers focused on the elements of effective instruction embedded in the walkthroughs and believed improvements in instruction and student achievement would result. Not all feedback on walkthroughs was positive, as one study reported an increase in the anxiety levels of teachers during the visits (Valli & Buese, 2007).

Preparing for walkthroughs. School leaders and teachers need to be adequately prepared before walkthroughs are introduced, and clear guidelines developed for all participants (Graf & Werlinich, 2002). Teachers should know the purpose of the walkthroughs and the observer's expectations (Lawler, 1991). Leaders should anticipate and develop strategies to defuse anxiety. The length and frequency of the walks should be shared, as well as making explicit what information is being gathered and how it will be used. Other areas for consideration with regards to walkthroughs include: deciding whether they will be announced or unannounced visits (Pieczura [2012] suggested that unplanned walkthroughs reveal more information regarding actual everyday practices within classrooms), providing training for observers and teachers, and deciding how the information obtained from the walkthrough will be used in the teacher's evaluation (Kachur et al., 2010).

Impacts of walkthroughs. Although research on "walkthroughs is limited in terms of demonstrating a direct cause-and-effect relationship between the tool and school, teacher, and student improvement" (Kachur et al., 2010, p. 25), a positive correlation was found when combining walkthroughs with other practices (Kachur et al., 2010). The tool increased the principal's awareness of classroom practice and helped leaders plan for professional development, which should lead to improvement in student achievement (Kachur et al., 2010). Pitler and Goodwin (2009) advised that walkthroughs should be used to share best practices and identify opportunities for growth, but should not be used for teacher evaluation.

Portfolios

Reflection is a deliberate process that can lead to cognitive growth if adequate time is provided for a focus on learning (Attinello, Lare, & Waters, 2006; Wade & Yarbrough, 1996). Models of professional development that encourage teachers to "reflect critically on their daily practices…enhance their capacity to understand complex subject matters from the perspectives of diverse learners" (Xu, 2003, p. 348). Portfolios of professional practices inherently and deliberately incorporate some degree of reflection, may positively impact the professional culture of a school (Attinello et al., 2006), and "offer both improved evaluation design elements and greater teacher involvement" (Tucker, Stronge, Gareis, & Beers, 2003, p. 574). If the criteria for portfolio development is aligned with desired educational outcomes, and the teacher is committed to honest reflections regarding practice, a portfolio can impact the professional growth of the teacher (Blake, Bachman, Frys, Holbert, Ivan, & Sellitto, 1995). Growth is maximized if professional dialogue occurs about the relationship between portfolio contents and the standards they represent (Gelfer, Xu, & Perkins, 2004; Riggs & Sandlin, 2000).

A portfolio compiled by teachers can serve several functions: as a formative assessment, to illuminate areas of strength and weaknesses of professional educators; as a summative assessment, if included in the teachers' formal evaluation process; and as a self-assessment, providing teachers the opportunity for focused reflection on their teaching practices (Berrill & Whalen, 2007; Riggs & Sandlin, 2000). Two of the roles for educational portfolios may seem contradictory: honest and open reflections on one's weaknesses for formative assessment purposes are unlikely to occur, and are unreasonable to expect if they are included in a summative assessment (Berrill & Whalen, 2007; Centra, 2000; Peterson, Stevens, & Mack, 2001; Riggs & Sandlin, 2000). However, even the use of portfolios for summative evaluation fosters self-reflection about possible changes in practice (Knapper & Wright, 2001).

Portfolios provide administrators the opportunity to look closely at a practice as it unfolds over time, unlike the brief snapshots available during a single observation; in addition, the portfolio process encourages the "reflection on those variables not easily captured during classroom observation" (Riggs & Sandlin, 2000, p. 24). Portfolios enhance the "documentation of assessment and professionalism" giving administrators a broader view of these important teacher qualities, separate from classroom observations (Tucker et al., 2003).

In the process of selecting evidence aligned to a standards-based portfolio, teachers may be alerted to areas of weakness if they are unable to find evidence to the contrary (Gelfer et al., 2004). "Ideally, this realization promotes the teacher's own pursuit of professional development in weak areas" (Riggs & Sandlin, 2000, p. 25). If artifacts are accompanied with explanations on their relationship to teaching, administrators gain deeper insight into the teacher's practices (Tucker et al., 2003).

Research. Studies examining the use of teacher portfolios have been mainly limited to the context of teacher preparation programs; little research in the use of portfolios for professional development of practicing teachers (Berrill & Whalen, 2007) and teacher evaluation (Xu, 2003) has been conducted. In the 1980s, the Utah Teacher

Evaluation Project attempted to incorporate the use of teacher portfolios to inform summative ratings (Peterson et al., 2001). A small pilot of this system revealed serious problems with using portfolios for this sole purpose including: (1) while portfolios may highlight excellence in teaching practices, the lack of uniformity in the portfolio structure makes it difficult to make fair comparisons, and (2) many personal qualities desired in a teacher (e.g., persistence, inspiration, personal interactions with students) are not readily observed through the portfolio collection process (Peterson et al., 2001, pp. 125 – 127).

In another study, portfolios were used to evaluate teacher performance under a newly implemented, performance-based compensation plan in Douglas County, Colorado (Wolf, Lichtenstein, Bartlett, & Hartman, 1996). Despite the use of volunteers, who were more likely to support the pilot, the teachers acknowledged that the process "encouraged them to clarify their instructional goals and more closely examine their teaching practices" (Wolf et al., 1996, p. 285). Administrators remarked that the examination of teachers' portfolios provided them with insights regarding the teachers' classroom practices and instructional philosophies (Wolf et al., 1996).

In a case study examining the impact of portfolios for collegial reflection on a specific area of interest, a small group of elementary teachers used portfolios over a two-year period (Xu, 2003). Positive impacts were noted in the following areas (Xu, 2003):

1. Professional learning: regardless of the level of experience, teachers reported the project enabled them "to approach their work more meaningfully and purposefully" (p. 352); and teachers revealed the project helped them to better know their students and meet their needs.

2. Professional collaboration: the project became a vehicle for connecting new teachers with more experienced teachers; teachers reported the process changed their working relationship with administrators; and “some teachers started to view themselves as agents of systematic change” (p. 354).

Several significant findings have been reported in studies investigating teachers’ opinions of portfolios. With regard to professional growth, Tucker et al. (2003) stated: “teachers reported more self-reflection as a result of portfolios but the self-reflection had little impact on teaching practice” (p. 591). Tucker et al. (2003) suggested additional mechanisms be included to help teachers connect their work on portfolios with activities that impact instructional practice. In another study, Attinello et al. (2006) reported an interesting comment by one teacher: “I could put together a really nice portfolio and not be a very good teacher. Conversely, a great teacher might not create a good portfolio” (p. 141). Additionally, both teachers and administrators agree that portfolios can provide a more comprehensive view of teacher performance, but caution that they may not be an accurate reflection of what actually occurs in the classroom (Attinello et al., 2006).

Research regarding assessment of the contents of portfolios has been scant, and limited information on the reliability and validity of evaluators’ ratings exists (Centra, 2000; Tucker et al., 2003). Portfolios that represent a comprehensive picture of teaching are believed to have face validity (Knapper & Wright, 2001), and more accurate and comprehensive than a traditional classroom observation (Attinello et al., 2006). In a multi-year study on the use of portfolios in a county-wide school district in Virginia, Tucker et al. (2003) found 90% of artifacts included by teachers had content validity. Concerns with accuracy can be addressed by administrators conducting regular classroom observations, looking for evidence to support portfolio presentations (Attinello et al., 2006).

Consistency in portfolio ratings is important. Reliability of portfolio assessments may be affected by “subjective impressions and personal relationships between the rater and the teacher assessed” (Van der Schaff, Stolling & Verloop, 2005, p. 47). Another investigation into the use of portfolios for program evaluation reported some difficulty with inter-rater reliability, however, when scoring open-ended tasks (Johnson, McDaniel, & Willeke, 2000). Some relevant findings of this study include (Johnson et al., 2000):

1. Training of raters on the types of evidence relevant to the purpose of the portfolio may improve rater reliability.
2. Based on reliability theory, composite scores are more reliable than subtest scores; this suggests the use of summative scores across several dimensions and/or by multiple raters improve reliability.
3. Evaluations of individual components of a portfolio are more reliable if raters are provided with specific “look-fors” (p. 78) regarding the proffered evidence or artifacts.

An important question regarding administrators’ abilities to distinguish among levels of teaching performance when assessing portfolios was investigated (Tucker et al., 2003). When the final ratings produced with portfolio evaluations were compared to prior evaluations based on traditional observations alone, a much greater differentiation occurred (Tucker et al., 2003). It was suggested that administrators are better able to recognize differences in teacher performance with the additional insight into instruction provided by portfolios (Tucker et al., 2003).

Recommendations. Portfolios provide authentic views of the complex process of teaching, and promote “active involvement of participants, encouragement of refection and self-assessment, and facilitation of collaborative interaction” (Tucker et al., 2003, p. 575). Portfolios can “empower teachers to take charge and have a more active voice in their evaluation” (Attinello et al., 2006, p. 134). However, questions remain on the effectiveness of portfolios for teacher evaluation. To be relevant, portfolio evaluation must be based on specific criteria and aligned with particular standards and important classroom practices (Riggs & Sandlin, 2000) to prevent a miscellaneous collection of artifacts that have no “relationship to critical thinking or teacher reflection” (Blake et al., 1995, p. 44). Providing teachers with a model of an exemplar portfolio can assist them with selection of artifacts and evidence representing key concepts (Moore & Bond, 2002). Administrators can support teachers by providing them ongoing feedback during the process (Moore & Bond, 2002) and adequate time to develop and reflect on portfolio contents (Attinello et al., 2006; Tucker et al., 2003).

Summary

In the Commonwealth of Pennsylvania, legislation mandating a new, high-stakes teacher evaluation process went into effect for the 2013-14 school year. School districts must implement Charlotte Danielson’s *Framework for Teaching* with all classroom teachers (unless a bargaining unit contract specifies another evaluation process). While the requirement to use Danielson’s clinical observation model with a portion of classroom teachers, use of an approved differentiated supervision model is required for the rest. A unique opportunity exists to compare the effects of the two models on important educational outcomes. The purpose of this study is to explore the impact of the new PA state-mandated, high-stakes teacher evaluation processes on the use of classroom instructional practices by teacher participants.

The proposed research study will use a three-pronged approach to explore the impact of the two evaluation models on teachers' (a) classroom instructional practices and (b) beliefs regarding self-efficacy:

1. All teachers will be rated on their instructional practices in Domain 2 and Domain 3 of Danielson's framework by the researcher. Each teacher will be observed twice for an entire classroom period, once in the fall and again in the spring.
2. An administrator will evaluate each classroom teacher using one of the two models permitted in the PA legislation. Summative evaluation ratings will be collected from the principals at the close of the school year.
3. All teachers will be given the opportunity to voluntarily and anonymously complete the Teachers' Self-Efficacy Survey, once at the beginning and at the end of the school year.

A tremendous amount of human and financial resources have been expended to develop teacher evaluation protocols to meet demands for accountability. The potential for these protocols to impact teachers' use of classroom best practices is an important consideration for the educational community as well as for policy-makers. While the national and state focus is on teacher accountability and complex systems to evaluate effective teaching, unless the evaluation process eventually leads to improved teaching practices, improved student learning may not result.

Chapter 3

Methodology

The focus school district is a small, rural, K-12 public school system in Western Pennsylvania with a student population of 1,504. There are a total of 111 classroom teachers in two buildings, a K-6 elementary and a 7-12 junior-senior high school, on the same campus. There are four building-level administrators in the district, one principal and one assistant principal in each building. All four administrators assumed their present positions in July 2010.

In order to meet the state's guidelines for rotating teachers through the two evaluation protocols (Formal Observation Model and Differentiated Supervision Model) over three to four years, approximately one-third of teachers are placed in the Formal Observation Model and two-thirds in the Differentiated Supervision Model. Due to the extensive time required to complete a Formal Observation, an attempt was made to equalize the numbers of teachers in each evaluation model to balance administrators' responsibilities. Elementary principals completed 18 Formal Observations and 37 Differentiated Supervision evaluations, and high school principals completed 17 Formal Observation and 39 Differentiated Supervision evaluations.

A number of research studies have been completed on the use of a standards-based teacher evaluation model such as Danielson's *Framework for Teaching* to determine teacher effectiveness. No studies have been reported that determine if there is a relationship between teacher participation in a standards-based evaluation model and changes in teachers' classroom instructional practices. The purpose of the proposed research is to examine the relationship between these variables.

Participants

Study setting. In its broadest scope, this study is intended to address the population of teachers participating in the newly state-mandated, standards-based evaluation model of teacher effectiveness in Pennsylvania. In order to control for variability across different districts in terms of training of teachers and evaluators in the use of a standards-based evaluation model, stage of model implementation, demographics of student populations, and contractual limitations on the use of teacher evaluation data, this study will be conducted in one rural school district in western Pennsylvania.

Population and sampling plan. There are 111 classroom teachers in grades K-12 in the target district. Thirty-five teachers will be assigned to the Formal Observation Model of the Danielson framework and 76 teachers to the Differentiated Supervision Model. Four building principals and two central office administrators will supervise the evaluation of small groups of classroom teachers from both groups.

Instrumentation

Four instruments will be used to collect data in this research study:

(1) Classroom observations of the instructional practices of all teachers will be based on Danielson’s *Framework for Teaching* rubric for Domains 2 and 3. The entire instrument is shown in Appendix A. Each teacher will be evaluated twice with this tool, once during the first nine-weeks of the school year and again during the fourth nine-weeks. Domain 2 (Classroom Environment) contains five components and 14 elements, as shown in Table 1. The researcher will assign a numerical rating during the observation for each component (0 = Failing, 1 = Needs Improvement, 2 = Proficient, 3 = Distinguished), based on evidence of the presence of critical attributes relevant for each proficiency level (shown in Appendix B).

Table 1

Components and Elements of Domain 2: The Classroom Environment

Component	Elements
A. Creating an Environment of Respect and Rapport	1. Teacher interactions with students 2. Student interactions with other students
B. Establishing a Culture for Learning	1. Importance of the content and of learning 2. Expectations for learning and achievement 3. Student pride in work
C. Managing Classroom Procedures	1. Management of instructional groups 2. Management of transitions 3. Management of materials and supplies 4. Performance of non-instructional duties
D. Managing Student Behavior	1. Expectations 2. Monitoring of student behavior 3. Response to student misbehavior
E. Organizing Physical Space	1. Safety and accessibility 2. Arrangement and use of physical resources

Domain 3 (Instruction) consists of five components and 18 elements (see Table 2). The rating for each component is an average of the ratings of the elements within the component. The rating for the Domain is an average of the calculated ratings of the five components within the domain.

Table 2

Components and Elements of Domain 3: Instruction

Component	Elements
A. Communicating with Students	1. Expectation for learning
	2. Directions and procedures
	3. Explanations of content
	4. Use of oral and written language
B. Questioning and Discussion Techniques	1. Quality of questions/prompts
	2. Discussion techniques
	3. Student participation
C. Engaging Students in Learning	1. Activities and assignments
	2. Grouping of students
	3. Instructional materials and resources
	4. Structure and pacing
D. Using Assessment in Instruction	1. Assessment criteria
	2. Monitoring of student learning
	3. Feedback to students
	4. Student self-assessment and monitoring of progress
E. Demonstrating Flexibility and Responsiveness	1. Lesson adjustment
	2. Response to students
	3. Persistence

During the 2012-13 school year, all administrators/supervisors in the district received a proficient rating in the use of Danielson's *Framework for Teaching.* The 30-hour series of training modules were provided by the Pennsylvania Department of Education's licensing agreement with Teachscape©. A copy of the researcher's certificate for proficiency is shown in Appendix C.

(2) Teacher self-efficacy ratings will be gathered through voluntary and anonymous online surveys. All teachers will have the opportunity to complete the survey at the beginning and again at the end of the school year. The long-form version of the Teachers' Sense of Efficacy Scale (see Appendix D), developed by Tschannen-Moran and Hoy (2001), will be used to generate teacher self-efficacy beliefs regarding student engagement, instructional strategies, and classroom management practices.

(3) Summative evaluation data for 35 classroom teachers will be generated by principals/supervisors during the implementation of the Formal Observation Model, aligned to the Danielson framework and protocol. At the end of the school year, principals/supervisors will provide two ratings for each teacher, one for Domain 2 and one for Domain 3.

(4) The school district in this study selected the Portfolio Mode for its implementation of the Differentiated Supervision Model, in which teachers examine their own practice and share artifacts or evidence of their performance level in each domain of the Danielson framework (PDE, 2013). District administrators developed criteria upon which the portfolio will be evaluated (Appendix E). Summative evaluation data for this study of the 76 classroom teachers in the Portfolio Mode will be generated by

principals/supervisors during the teachers' portfolio presentations of artifacts/evidence on components in Domains 2 and 3.

Independent variables. Classroom teachers will participate in one of the two models mandated by the Commonwealth's new evaluation protocol: the Formal Observation Model or the Differentiated Supervision Model, both based on Charlotte Danielson's *Framework for Teaching.* Summative educator effectiveness ratings collected by the principals/supervisors will be used to measure these variables.

Dependent variable. Classroom observations conducted by the researcher will measure the dependent variable, the ratings of teachers' use of classroom instructional practices, aligned to Charlotte Danielson's *Framework for Teaching*.

Moderating variables. While the primary goal of this study is to compare observation ratings of classroom practices of the two groups of teachers participating in different evaluation protocols, other relationships relevant to the evaluation process will be explored. The summative educator effectiveness ratings in Domains 2 and 3completed by principals/supervisors will be compared to the final observation ratings of classroom practices conducted by the researcher. The results to teachers' self-efficacy ratings will be compared to classroom observation ratings and to summative educator effectiveness ratings.

Analyses of classroom instructional practices. A sample of the rubric for one element of Domain 2 is provided in Table 3.

Table 3

Sample Rubric for Domain 2A, Element 1, Teacher Interactions with Students

Rating	Descriptive Behaviors
0 - Failing	Teacher interaction with at least some students is negative, demeaning, sarcastic, or inappropriate to the age or culture of the students. Students exhibit disrespect for the teacher.
1 - Needs Improvement	Teacher-student interactions are generally appropriate but may reflect occasional inconsistencies, favoritism, or disregard for students' cultures. Students exhibit only minimal respect for the teacher.
2 - Proficient	Teacher-student interactions are friendly and demonstrate general caring and respect. Such interactions are appropriate to the age and cultures of the students. Student exhibit respect for the teacher.
3 - Distinguished	Teacher interactions with students reflect genuine respect and caring for individuals as well as groups of students. Students appear to trust the teacher with sensitive information.

Teacher sense of self-efficacy surveys. The long form of the Teacher Sense of Self-Efficacy Survey consists of 24 questions, eight for each of three subscales. Responses to each question were rated on a 9-point Likert scale in which teachers were asked how much they can do in various situations. Choices ranged from 1 (Nothing) to 9 (A Great Deal). Sample questions for each subscale are shown in Table 4.

Table 4

Sample Self-Efficacy Questions

Efficacy Subscale	Questions
Student Engagement	1. How much can you do to get through to the most difficult students? 2. How much can you do to help your students think critically? 3. How much can you do to motivate students who show low interest in school work?
Instructional Strategies	1. How much can you gauge student comprehension of what you have taught? 2. How much can you do to adjust your lessons to the proper level for individual students? 3. How much can you use a variety of assessment strategies?
Classroom Management	1. How much can you do to control disruptive behavior in the classroom? 2. To what extent can you make your expectations clear about student behavior? 3. How well can you establish routines to keep activities running smoothly?

Reliability and validity of the teachers' sense of efficacy scale. In a Scree Test performed on the 36-item teacher self-efficacy instrument developed by Tschannen-Moran and Hoy (2001), three factors were extracted: (1) efficacy for instructional strategies, (2) efficacy for classroom management, and (3) efficacy for student engagement. Using the eight items with the highest loadings on each factor, a 24-item instrument was produced, with loadings ranging from 0.50 to 0.78 and intercorrelations between the three subscales were 0.60, 0.70, and 0.58, respectively ($p < 0.001$) (Tschannen-Moran & Hoy, 2001, p. 799). Additional descriptive statistics for the three factors are shown in Table 5.

Table 5

*Descriptive Statistics for Subscales of Teacher Efficacy Scale**

	Mean	SD	α	Eigen value	Cum %
Instructional Strategies	7.3	1.1	0.91	10.38	43.25
Classroom Management	6.7	1.1	0.90	2.03	51.72
Student Engagement	7.3	1.1	0.87	1.62	58.47

*Tschannen-Moran & Hoy, 2001, p. 800

In a follow-up study of the scale with in-service teachers ($N = 111$), the three subscale factors accounted for 54% of the variance in the teachers' responses (Tschannen-Moran & Hoy, 2001, p. 799). The reliability of the 24-item scale was 0.94, indicating the total score and the subscale scores are reliable measures of efficacy (Tschannen-Moran & Hoy, 2001, p. 801). Evidence of construct validity was determined as a result of a positive correlation of this scale with other measures of personal teaching efficacy; in addition, this scale was found to "capture a wider range of teaching tasks" (Tschannen-Moran & Hoy, 2001, p. 801).

Fives and Buehl (2009) examined the factor structure of the Teachers' Sense of Efficacy Scale and determined the "three-factor conceptualization of teacher efficacy appears to be appropriate for practicing teachers" (Fives & Buehl, 2009, p. 132). Research examining the psychometric properties of the Teachers' Sense of Efficacy Scale found the same three distinct factors as those presented by Tschannen-Moran and Hoy (2001), with comparable scale reliabilities, intercorrelations, means, and standard deviations; in addition, the factor structure held for teachers at the elementary, middle, and secondary levels (Heneman, Kimball, & Milanowski, 2006). Heneman et al. (2006) concluded the Teachers' Sense of Efficacy Scale should be the preferred measure of teacher efficacy due to "its replicable psychometric properties, behavioral richness in capturing the teacher role, and predictive capacity for explaining significant variance in teacher classroom performance" (p. 13).

Procedures

Teacher self-efficacy surveys and classroom observations of teachers' instructional practices will be completed in the same time frame (each will be given twice to every teacher during the first and fourth nine-weeks of the school year). In January, administrators/supervisors will begin the evaluations of teachers in the Formal Observation Model. Teachers in the Differentiated Supervision Model will be afforded time during scheduled in-services to gather artifacts and evidence for their individual portfolios. Portfolio presentations will be held during the last week of school.

Proposed Data Analysis

A two-group, pretest-posttest, quasi-experiment will be used to determine the relationship between teacher participation in one of two evaluation models and ratings of their classroom instructional practices taken before and after their participation in the evaluation protocol. Teachers participating in the Formal Observation Model will be compared with teachers participating in the Differentiated Supervision Model.

In order to address the primary hypothesis, "there is a relationship between teacher participation in the Formal Observation Model and implementation of best practices in classroom instruction", an independent samples *t*-test will be used to determine whether there is a statistical difference between the means of the Classroom Instructional Practices' scores of the teachers in each evaluation protocol (Formal Observation vs.. Differentiated Supervision) collected during Observation 1 and Observation 2. A paired-samples *t*-test will be used to determine if there is a statistical difference between the means of the Classroom Instructional Practices' scores from Observation 1 compared to Observation 2. A General Linear Model Repeated-Measures Analysis of Variance will be conducted to compare the means generated by participants in Observation 1 with Observation 2.

A secondary investigation will be conducted to explore possible relationships between the self-efficacy ratings of teachers in each evaluation protocol. An independent samples *t*-test will be used to determine whether there is a statistical difference between the self-efficacy ratings of teachers in each evaluation protocol (Formal Observation vs. Differentiated Supervision).

Finally, a possible relationship between the Classroom Instructional Practices' ratings and the summative Educator Effectiveness scores will be explored. A paired-samples *t*-test will be conducted on ratings collected during Observation 2 with Educator Effectiveness scores collected for all teachers. All hypotheses will be tested at a minimum .05 level of significance.

Limitations. The major weaknesses of this research protocol include various threats to validity and reliability, generalizability, and sample size. Social threats to internal reliability are possible since all participants are part of the same faculty. Generalizability of findings to other populations is limited by differences in various demographical and contextual factors of other populations. The power and effect size of the findings can be diminished by the small sample sizes in this study.

Inter-rater reliability threats may occur with Educator Effectiveness ratings, as they will be collected by different administrators/supervisors and use different instruments for the two evaluation protocols. Although administrators/supervisors have been trained and tested on their ability to discern among proficiency levels within the Danielson framework (as applied to teachers in the Formal Observation Model), the instrument applied to the teachers in the Differentiated Supervision Model was created by the administrators and as a result, may lack construct validity. Concurrent validity of this instrument may exist if the administrators are not able to distinguish between the four levels of proficiency.

Chapter 4

Results

As a result of the new state-mandated teacher evaluation system, classroom teachers must be cycled through one of two evaluation protocols beginning with the 2013-14 school year: a Formal Observation Model or a Differentiated Supervision Model. Each classroom teacher is expected to be evaluated in the Formal Observation Model once every three to five years. The current investigation examined the possible relationships between the type of evaluation protocol experienced by classroom teachers and their ratings in three different constructs.

The first construct, Classroom Instructional Practices was based on full-period classroom observations of all teachers, conducted by the researcher at the beginning and end of the school year. The second construct, Teacher Self-Efficacy ratings, were collected through anonymous and voluntary online surveys conducted at the beginning and the end of the school year. The third construct, Educator Effectiveness ratings, were determined by building and district administrators as required by the new legislation for teacher evaluation in PA. The legislation mandates teachers in the Formal Observation Model be evaluated by administrators using Danielson's *Framework for Teaching* rubric, while teachers in the Differentiated Supervision Model are to be rated through one of three modes: (1) Peer Coaching, (2) Action Research, or (3) Portfolio.

Demographics of the participants, disaggregated by type of evaluation protocol, are provided first. Descriptive and preliminary analysis of Classroom Instructional Practices, disaggregated by Domains and type of evaluation protocol are presented next. The following section reports the descriptive and preliminary analysis of the results to the teachers' self-efficacy survey responses, disaggregated by type of evaluation protocol and efficacy categories. The final section describes the descriptive and preliminary analysis of Educator Effectiveness ratings, disaggregated by Domain and evaluation protocol.

Teacher Evaluation Models

Demographics. Four demographic categories of the classroom teaching staff are reported in Table 6 (gender, years of service in the district, school building, and type of evaluation protocol). There are approximately equal numbers of elementary (*n* = 55) and high school (*n* = 56) classroom teachers in the district. Thirty-two percent (*n* = 35) of classroom teachers were placed in the Formal Observation Model of the new teacher evaluation system and sixty-eight percent (*n* = 76) were evaluated with the Differentiated Supervision Model. Thirty percent (*n* = 33) of the teachers are male and seventy percent (*n* = 78) are female. Less experienced teachers (1-10 years of service) make up 45% (*n* = 50) of the sample population, while teachers with the most experience (more than 30 years) make up only 9% (*n* = 10) of the sample.

Table 6

Demographics of Sample Population by Evaluation Protocol

Demographic	Category	Formal Observation	Differentiated Supervision	Total Numbers
Building				
	Elementary (K-6)	18	37	55
	High School (7-12)	17	39	56
Gender				
	Male	7	26	33
	Female	28	50	78

Years in District				
	1-10	23	27	50
	11-20	7	26	33
	21-30	5	13	18
	30+	0	10	10
Tenure Status				
	Tenured	22	76	98
	Non-Tenured	13	0	13
Totals		35	76	111

Construct 1: Classroom Instructional Practices

Full period observations of each classroom teacher were conducted twice by the researcher, once during the first nine-weeks of the 2013-14 school year and during the fourth nine-weeks, using the Danielson (2011) rubric for Domains 2 and 3 (Appendix A). Each observation was conducted by the same researcher, lasted a minimum of 30 minutes, and resulted in a total of 222 observations. A majority of the observed classes (72%, *n* = 159) were core content courses: English Language Arts, Mathematics, Science, and Social Studies as indicated in Table 7.

Table 7

Subjects Taught during Classroom Observations Conducted by Researcher

Subject	Round 1	Round 2	Total
English Language Arts	33	30	63
Mathematics	28	24	52
Science	12	11	23
Social Studies	9	12	21
Business/Vocational	7	9	16
Art/Music/PE	12	11	23
Other	10	14	24

Classroom instructional practices may vary according to the needs of the students in the classroom. All types of classrooms were observed in order to capture a wide range of instructional and classroom management strategies. These values are listed in Table 8.

Table 8

Category of Classrooms Observed by Researcher

Subject	Observation 1	Observation 2	Totals
Regular Education	89	87	176
Special Education	11	12	23
Inclusion/Basic	4	1	5
Honors	7	11	18

The vast majority of classes (79%, *n* = 176) were regular education, consisting of students with varying ability levels. Inclusion/basic classes were taught by a regular education teacher, but contained students with special needs who were aided by the presence of a learning-support educator. Special education classes were taught by certified special education teachers and included only students with special needs, ranging from mild learning disabilities to autistic or emotional support. Honors level classes typically carry college level or Advanced Placement credit.

Descriptive Analysis of Classroom Instructional Practices' Domains.

Descriptive analyses were conducted on the changes in the ratings for individual participants, comparing Observation 2 to Observation 1 (see output in Appendix F). Summary statistics are shown in Table 9.

Table 9

Comparison of Overall Observation Ratings by Evaluation Protocol

	Obs. 1	Obs. 2
Formal Observation Participants	2.14	2.19
Differentiated Supervision Participants	**2.10**	**2.19**

Descriptive analyses of the domains were performed to assess the assumptions of normality. Summary statistics are found in Table 10 (output provided in Appendix G).

Table 10

Summary Statistics

Ratings	N	M	SD	Skewness	Kurtosis
Observation 1: Domain 2					
Formal Observation	35	2.13	.34	-.93	2.41
Differentiated Supervision	76	2.10	.29	-.62	.96
Observation 1: Domain 3					
Formal Observation	35	2.16	.32	-1.94	5.06
Differentiated Supervision	76	2.11	.35	-1.12	1.28
Observation 2: Domain 2					
Formal Observation	35	2.13	.19	.09	.10
Differentiated Supervision	76	2.15	.20	-.35	.47
Observation 2: Domain 3					
Formal Observation	35	2.25	.16	.08	-.84
Differentiated Supervision	76	2.22	.23	-.94	1.90

Reliability analyses. High reliabilities (based on Tabachnick & Fidell, 2013 guidelines) were revealed among the five components of Domains 2 and 3 in all Observation 1 ratings (Cronbach's α between .69 and .86). However, lower reliabilities occurred with the components of the two Domains in all Observation 2 ratings (Cronbach's α between .56 and .67). See Table 11 and output in Appendix H.

Table 11

Reliability of Observation Ratings

Subscale	Cronbach's α
Observation 1	
Domain 2	.78
Formal Observation	.84
Differentiated Supervision	.74
Domain 3	.83
Formal Observation	.69
Differentiated Supervision	.86
Observation 2	
Domain 2	.65
Formal Observation	.67
Differentiated Supervision	.64
Domain 3	.60
Formal Observation	.56
Differentiated Supervision	.61

Assumptions of normality. In a normal distribution, the values of skewness and kurtosis should approach zero. Since all variables have a skewness value less than |2.0|, the assumption of normality is tenable for all variables (Tabachnick & Fidell, 2013). The kurtosis values are all less than |7.0|, indicating the assumption of normality is tenable for all variables (Tabachnick & Fidell, 2013).

The Kolmogorov-Smirnov (K-S) test of normality supports the significance of normal distributions in the ratings for participants in the Formal Observation Model for Domain 2 of Observation 1, *D*(35) = .12, *p* = .20, Domain 2 of Observation 2, *D*(35) = .10, *p* = .20, and Domain 3 of Observation 2, *D*(35) = .13, *p* = .17. For participants in the Differentiated Supervision Model, K-S tests of normality supports the significance of normal distributions in the ratings for Domain 3 of Observation 2, *D*(76) = .08, *p* = .20 (see output in Appendix I).

The K-S tests of normality were significantly non-normal for participants in the Formal Observation Model for Domain 3 of Observation 1, *D*(35) = .17, *p* = .01, as were the ratings for participants in the Differentiated Supervision Model for Domain 2 of Observation 1, *D*(76) = .10, *p* = .04, Domain 3 of Observation 1, *D*(76) = .14, *p* = .001, and Domain 2 of Observation 2, *D*(76) = .11, *p* = .02 (see output in Appendix I). Since the skewness and kurtosis of these variables are within the acceptable ranges, these significant 1 Sample K-S results are not concerning. The 1 Sample K-S test is sensitive to sample sizes in excess of *n* = 100, since it is based on a chi-square distribution (Tabachnick & Fidell, 2013). Therefore, normality of these variables is assumed tenable.

Homogeneity of variance. Levene's test of homogeneity of variance supports the assumption that the variances are not significantly different (see output in Appendix I). For Observation 1 ratings: Domain 2, *F*(1,109) = 1.06, *p* = .31 and Domain 3, *F*(1,109) = .93, *p* = .34; for Observation 2 ratings: Domain 2, *F*(1,109) = .02, *p* = .88 and Domain 3, *F*(1,109) = 2.22, *p* = .14.

Correlations. Correlational analyses reveal significant positive relationships between individual domains and overall ratings in both rounds of observations, as shown in Table 12 (output in Appendix J).

Table 12

Correlations of Ratings for Domains and Overall Ratings (N = 111)

Dependent Variables Compared	r
Observation 1	
Domain 2 vs. Domain 3	.73**
Domain 2 vs. Overall Rating	.92**
Domain 3 vs. Overall Rating	.94**
Observation 2	
Domain 2 vs. Domain 3	.70**
Domain 2 vs. Overall Rating	.92**
Domain 3 vs. Overall Rating	.93**
Observation 1 vs. Observation 2	.32**

** Correlation is significant at the .01 level (two-tailed).

A second correlational analysis was conducted on the data, separated by Type of Evaluation Protocol. Significant positive relationships are reported in Table 13 (see output in Appendix K).

Table 13

Correlations of Domains 2 and 3 by Evaluation Protocol

Dependent Variables Compared	n	r
Formal Observation Protocol Participants (*n* = 35)		
Observation 1	35	.84**
Observation 2	35	.71**
Differentiated Supervision Protocol Participants		
Observation 1	76	.69**
Observation 2	76	.72**

** Correlation is significant at the .001 level (two-tailed).

Correlations between the initial and final overall observation ratings of participants in the Formal Observation Protocol were not significant, whereas the correlation for participants in the Differentiated Supervision Model were significant (see Table 14 and output in Appendix K).

Table 14

Correlations of Initial and Final Observations

Dependent Variables Compared	r	p
Formal Observation Protocol Participants (*n* = 35)		
Observation 1 vs. Observation 2	.23	.18
Differentiated Supervision Protocol Participants (*n* = 76)		
Observation 1 vs. Observation 2	.36**	.001

** Correlation is significant at the 0.01 level (two-tailed).

Independent t-tests. When two groups of participants are exposed to different treatments, independent (between-group) analyses may be used to compare the means of a measurement conducted on the groups. Since the two groups consist of different participants, mean ratings may differ because of participants' individual differences and not because of the treatment. An independent *t*-test can be used to examine the significance of any difference in mean ratings.

In this study, the Classroom Instructional Practices of teachers in two different evaluation protocols were measured using a four-point rubric based on the Danielson *Framework for Teaching*. Although the framework has four domains, only the practices of Domain 2 (Classroom Environment) and Domain 3 (Instruction) are observable during a classroom observation. During the first round of observations, participants in the Formal Observation Model had slightly higher ratings in each domain than participants in the Differentiated Supervision Model, as indicated in Table 15. These differences were not significant (see output in Appendix L). During the second round of observations, there were no significant differences between the average ratings in the two groups, although Formal Observation Model participants had slightly lower ratings in Domain 2 and slightly higher ratings in Domain 3.

Table 15

Independent Samples t-Test

	N	M	SE	*t* - statistic
Observation 1: Domain 2 Rating				
Formal Observation	35	2.13	.06	*t*(109) = .54, *p* = .59
Differentiated Supervision	76	2.10	.03	
Observation 1: Domain 3 Rating				
Formal Observation	35	2.16	.05	*t*(109) = .67, *p* = .50
Differentiated Supervision	76	2.11	.04	
Observation 2: Domain 2 Rating				
Formal Observation	35	2.13	.03	*t*(109) = -.45, *p* = .65
Differentiated Supervision	76	2.15	.02	
Observation 2: Domain 3 Rating				
Formal Observation	35	2.25	.03	*t*(109) = .62, *p* = .54
Differentiated Supervision	76	2.22	.03	

In conclusion, the results of the independent *t*-tests indicate there is no significant difference in the mean ratings of the two groups of participants.

Paired-samples t-tests. In order to compare changes in the means of two measurements collected from the same participants, a dependent (paired-samples) *t*-test is used. Initial and final average ratings on the Classroom Instructional Practices' rubric of the participants in each evaluation protocol for Domain 2 (Classroom Environment) and Domain 3 (Instruction) are found in Table 16 (see output in Appendix M).

Table 16

Paired-Samples t-Tests for Domains 2 and 3

Variables	M	SE	*t*-statistic	r
Domain 2 Ratings				
Formal Observation				
Initial	2.13	.06	*t*(34) = .08, *p* = .94	.01
Final	2.13	.03		
Differentiated Supervision				
Initial	2.10	.03	*t*(*75*) = 1.79, *p* = .08	.20
Final	2.15	.02		
Domain 3 Ratings				
Formal Observation				
Initial	2.16	.05	*t*(34) = 1.47, *p* = .15	.24
Final	2.25	.03		
Differentiated Supervision				
Initial	2.11	.04	*t*(75) = 2.46, *p* = .02	.27
Final	2.21	.03		

As seen in Table 16, the final ratings in Domain 2 and Domain 3 of participants in the Formal Observation Model were not significantly different than their initial observation ratings, whereas Domain 3 ratings of participants in the Differentiated Supervision Model were significantly greater (*p* = .02) in their final observations when compared to their initial observations, with a medium effect size (*r* = .27).

*GLM repeated measures analysis of variance (ANOVA).*When conducting several *t*-tests to compare pairs of groups, the probability of making a Type I error (falsely rejecting the null hypothesis) increases (Field, 2009). In order to compare the means generated when the same participants are rated before and after an experimental condition is applied, a repeated-measures ANOVA is appropriate. Multiple *t*-tests can result in positively biasing results; this bias can be eliminated with the use of a GLM repeated measures ANOVA. In this study, participants in each of the evaluation protocols were observed with the Classroom Instructional Practices' rubric at the beginning and end of the school year, and scores were calculated for the two domains of the rubric. A repeated-measures ANOVA considers both between-group and repeated measures. The repeated measures of the participants for each domain result in the two factors designated as the *Within-Subject* variables, as indicated in Table 17 (see output in Appendix N).

Table 17

Within-Subject Variables

Factor	Levels
1. Domain 2	1. Pre-Observation
	2. Post-Observation
2. Domain 3	1. Pre-Observation
	2. Post-Observation

The assumption of homogeneity of covariance structures was not tenable, but assumed not compromising to the interpretation of this data since the error *df* is greater than 20 (Tabachnick and Fidel, 2013). An accurate *F*-test in ANOVA is based on the assumption that two sets of scores are independent, generated by different participants (Field, 2009). The assumption of sphericity is used to assess the equality of variances of the differences between pairs of Classroom Instructional Practices' scores for each participant, and was found to be tenable.

The within-subjects analyses indicate that there is a significant difference in participants' scores from pretest to posttest for Domain 2 (Factor 1), from pretest to posttest for Domain 3 (Factor 2), and a significant difference in the scores for Domain 2 relative to Domain 3. The details of these results are presented in Table 18.

Table 18

Within-Subject Analysis

			Partial Eta	Observed
	F	Sig.	Squared	Power
Domain 2	**15.19**	**.00**	.12	.97
Domain 2 by Group	1.00	.32	.01	.17
Domain 3	**4.31**	**.04**	.04	.54
Domain 3 by Group	.36	.55	.00	.09
Domain 2 vs. Domain 3	**5.02**	**.03**	.04	.60

Figure 1 illustrates the pretest to posttest changes revealed for Domain 2 (Factor 1).

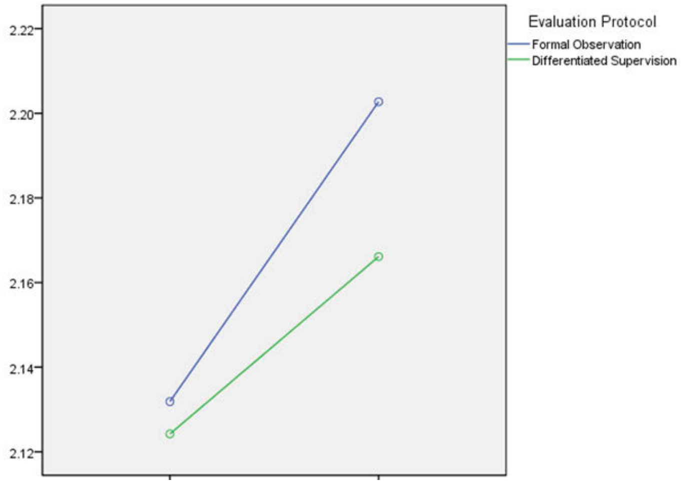


Figure 1. Illustration of Pretest to Posttest changes on Domain 2 across Protocols

As seen in Figure 1, the two groups began with similar pretest values; however, the Formal Observation Group shows great gains in change. As seen in Figure 2, the two groups differed more on Domain 3 (at pretest) relative to Domain 2. However, the two groups produced similar posttest scores. Figure 2 illustrates the pretest to posttest changes revealed for Domain 3 (Factor 2).

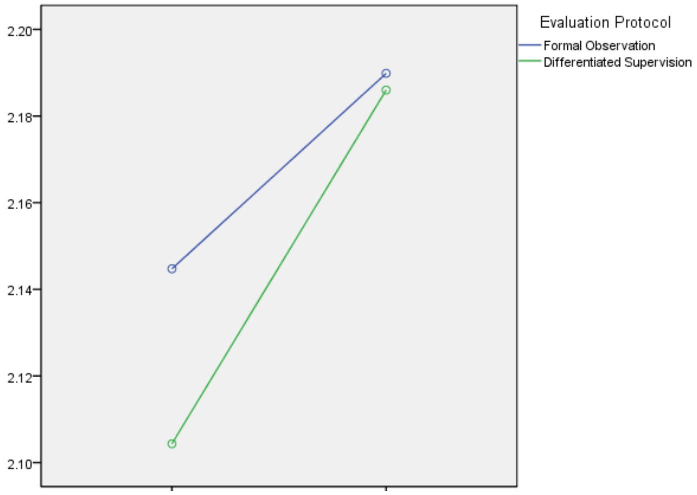


Figure 2. Illustration of Pretest to Posttest changes on Domain 3 across Protocols

Construct 2: Teacher Self-Efficacy

The long-form version of the Teachers' Sense of Efficacy Scale (see Appendix C), developed by Tschannen-Moran and Hoy (2001), was used to explore the relationships between teacher self-efficacy assessments of three subscales of the instrument (efficacy in student engagement, efficacy in instructional strategies, and efficacy in classroom management practices) and the type of evaluation protocol experienced by the participants. Subscale ratings from the initial survey are labeled as: Pre-Student Engagement, Pre-Instructional Strategies, and Pre-Classroom Management scores; those from the final survey are labeled as Post-Student Engagement, Post-Instructional Strategies, and Post-Classroom Management scores.

Composite survey results. There was a significant difference between the composite self-efficacy ratings (an average of the three subscales) of the participants in the two evaluation protocols measured at the beginning of the school year, as shown in Table 19 (output in Appendix O).

Table 19

Composite Self-Efficacy Ratings from Initial Survey

Evaluation Protocol	N	M	SE	*t*-statistic
Formal Observation	31	7.54	.68	*t*(82) = 2.33, *p* = .02
Differentiated	53	7.16	.76	

Subscale survey questions. All classroom teachers were sent the Teacher Self-Efficacy survey through Survey Monkey at the beginning and end of the school year. Responses were voluntary and anonymous, and the response rates are shown in Table 20.

A greater percentage of teachers in the Formal Observation protocol responded in both surveys. There was a decrease in all participation rates from the beginning to the end of the year.

Table 20

Teacher Self-Efficacy Response Rates

Evaluation Protocol	Initial Survey		Final Survey	
	n	Pct. of Total	n	Pct. of Total
Formal Observation	31	89%	26	74%
Differentiated Supervision	53	70%	49	64%
Totals	84	76%	75	68%

Analysis of self-efficacy ratings. Descriptive analyses of the variables were performed to assess the assumptions of normality. Table 21 provides the summary statistics for these variables (output provided in Appendix P).

Table 21

Summary Statistics

Variable	Evaluation Protocol	N	M	SD	Skewness	Kurtosis
Pre-Student Engagement						
	Formal Observation	31	7.17	.90	.19	-.57
	Differentiated Supervision	53	6.58	.99	-.29	-.35
Pre-Instructional Strategies						
	Formal Observation	31	7.64	.72	-.03	-.78
	Differentiated Supervision	53	7.41	.87	-.47	.28
Pre-Classroom Management						
	Formal Observation	31	7.83	.77	-.26	-.06
	Differentiated Supervision	53	7.49	.91	-.64	.51
Post-Student Engagement						
	Formal Observation	26	6.80	.72	.03	-.80
	Differentiated Supervision	49	6.54	1.05	.40	-.15
Post-Instructional Strategies						
	Formal Observation	26	7.27	.68	-.12	-.53
	Differentiated Supervision	49	7.42	.81	-.18	.39
Post-Classroom Management						
	Formal Observation	26	7.56	.80	-.96	1.47
	Differentiated Supervision	49	7.60	.90	-.39	-.46

Reliability analysis. All efficacy subscales collected in the initial and final teacher surveys had high reliabilities, between Cronbach's α = .82 and .86, as indicated in Table 22 (see output in Appendix Q).

Table 22

Reliability Coefficients

Subscale	Cronbach's α
Initial Survey Results	
Efficacy in Student Engagement	.85
Efficacy in Instructional Strategies	.83
Efficacy in Classroom Management	.86
Final Survey Results	
Efficacy in Student Engagement	.83
Efficacy in Instructional Strategies	.82
Efficacy in Classroom Management	.86

Assumptions of normality. In a normal distribution, the values of skewness and kurtosis should be zero. Since all variables have a skewness value less than |2.0|, the assumption of normality is tenable for all variables (Tabachnick & Fidell, 2013). The kurtosis values are all less than |7.0|, indicating the assumption of normality is tenable for all variables (Tabachnick & Fidell, 2013).

The Kolmogorov-Smirnov (K-S) test of normality supports the significance of normal distributions in all variables (see output in Appendix R). In the Formal Observation Model, the Pre-Student Engagement variable, $D(31) = .13$, $p = .20$, the Pre-

Instructional Strategies' variable, *D*(31) = .11, *p* = .20, and the Pre-Classroom Management variable, *D*(31) = .11, *p* = .20; at the end of the year, the Post-Student Engagement variable, *D*(26) = .10, *p* = .20, the Post-Instructional Strategies variable, *D*(31) = .10, *p* = .20, and the Post-Classroom Management variable, *D*(31) = .12, *p* = .20.

For the Differentiated Supervision Model, the Pre-Student Engagement variable, *D*(53) = .11, *p* = .20, the Pre-Instructional Strategies' variable, *D*(53) = .07, *p* = .20, and the Pre-Classroom Management variable, *D*(53) = .12, *p* = .07; the Post Student Engagement variable, *D*(49) = .09, *p* = .20, the Post Instructional Strategies' variable, *D*(49) = .10, *p* = .20, and the Post Classroom Management variable, *D*(49) = .11, *p* = .20.

Homogeneity of variance. Levene's test of homogeneity of variance supports the assumption that the variances are not significantly different (see output in Appendix S). For the Pre-Student Engagement variable, *F*(1,82) = .24, *p* = .63, the Pre-Instructional Strategies' variable, *F*(1,82) =.56, *p* = .46, and the Pre-Classroom Management variable, *F*(1,82) = .50, *p* = .48; for the Post-Student Engagement variable, *F*(1,73) = 2.37, *p* = .13, the Post-Instructional Strategies variable, *F*(1,73) = .71, *p* = .40, and the Post-Classroom Management variable, *F*(1,73) = .91, *p* = .34.

Correlations. There is a significant positive relationship between teachers' perceived efficacies in all three subscales (see output in Appendix T). Table 23 shows the correlations for all respondents in the aggregate.

Table 23

Correlation between Subscales of Teacher Self-Efficacy Surveys – All Participants

Subscale Comparisons	r
Initial Surveys	
Student Engagement vs. Instructional Strategies	.57**
Student Engagement vs. Classroom Management	.49**
Instructional Strategies vs. Classroom Management	.61**
Final Surveys	
Student Engagement vs. Instructional Strategies	.61**
Student Engagement vs. Classroom Management	.53**
Instructional Strategies vs. Classroom Management	.45**

The correlational analyses between subscales were also conducted by disaggregating the participants according to evaluation protocol. These results for participants in the Formal Observation Model appear in Table 24.

Table 24

Correlation between Subscales of Self-Efficacy Surveys of Formal Observation Teachers

Subscale Comparisons	r
Initial Surveys	
Student Engagement vs. Instructional Strategies	.68**
Student Engagement vs. Classroom Management	.56**
Instructional Strategies vs. Classroom Management	.50**
Final Surveys	
Student Engagement vs. Instructional Strategies	.44*
Student Engagement vs. Classroom Management	.66**
Instructional Strategies vs. Classroom Management	.47*

* Correlation is significant at the .05 level (two-tailed)
** Correlation is significant at the .01 level (two-tailed)

For participants in the Differentiated Supervision Model, correlation analyses are displayed in Table 25.

Table 25

Correlation between Subscales of Self-Efficacy Surveys of Differentiated Supervision Teachers

Subscale Comparisons	r
Initial Surveys	
Student Engagement vs. Instructional Strategies	.51**
Student Engagement vs. Classroom Management	.42**
Instructional Strategies vs. Classroom Management	.65**
Final Surveys	
Student Engagement vs. Instructional Strategies	.69**
Student Engagement vs. Classroom Management	.50**
Instructional Strategies vs. Classroom Management	.44**

** Correlation is significant at the 0.01 level (two-tailed)

Independent t-tests of pre/post self-efficacy ratings: In the first round of surveys, participants in the Formal Observation Model report a greater self-efficacy in all three subscales than participants in the Differentiated Supervision Model (output is shown in Appendix U). These results are presented in Table 26.

Table 26

Independent Samples Test for Self-Efficacy Ratings

Subscale	N	M	SE	*t* - statistic
Pre-Student Engagement				
Formal Observation	31	7.17	.16	*t*(82) = 2.73, *p* = .008
Differentiated Supervision	53	6.58	.14	
Pre-Instructional Strategies				
Formal Observation	31	7.64	.13	*t*(82) = 1.24, *p* = .22
Differentiated Supervision	53	7.41	.12	
Pre-Classroom Management				
Formal Observation	31	7.83	.14	*t*(82) = 1.74, *p* = .08
Differentiated Supervision	53	7.49	.12	
Post-Student Engagement				
Formal Observation	26	6.80	.14	*t*(73) = 1.12, *p* = .27
Differentiated Supervision	49	6.54	.15	
Post-Instructional Strategies				
Formal Observation	26	7.27	.13	*t*(73) = -.83, *p* = .41
Differentiated Supervision	49	7.42	.12	
Post-Classroom Management				
Formal Observation	26	7.56	.16	*t*(73) = -.20, *p* = .84
Differentiated Supervision	49	7.60	.13	

As indicated by the independent samples *t*-test, the only significant difference between the two groups is found on the pre-administration of the survey for the Student Engagement factor. In the post-administration of the survey, none of the subscale ratings had significant differences between the participants in the two evaluation protocols. Figure 3 provides a graphical depiction of these outcomes.

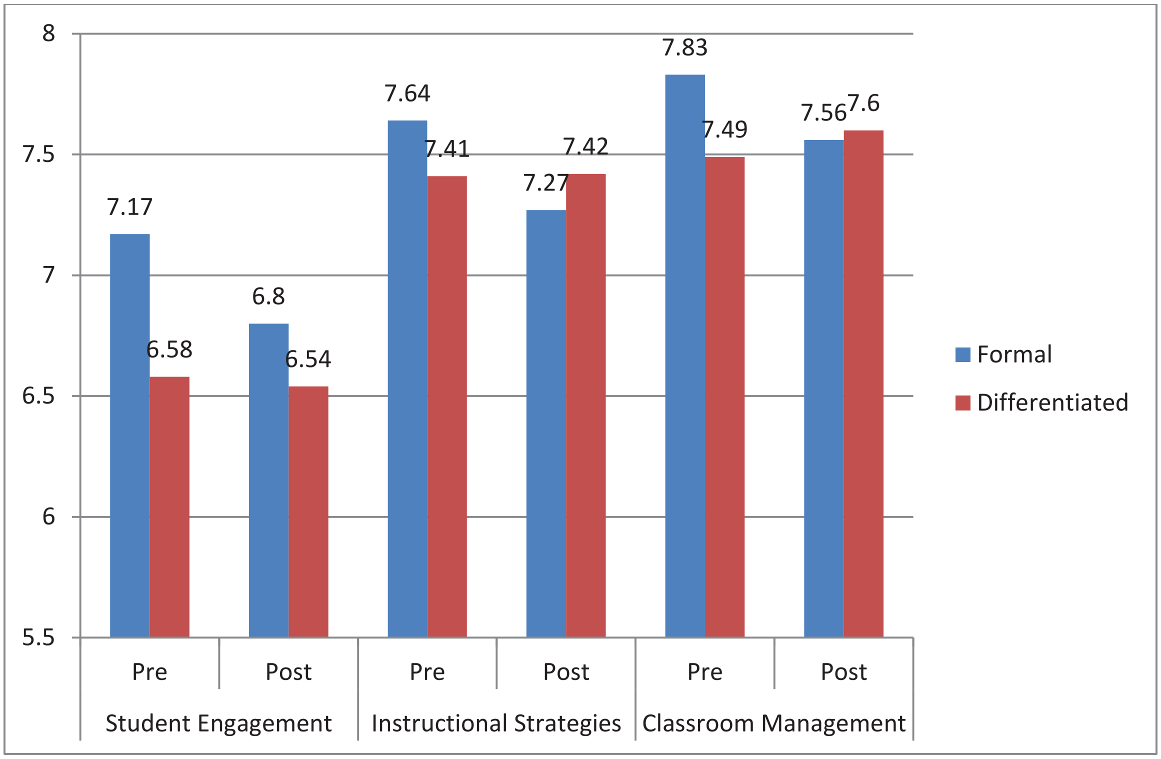


Figure 3. Pre- to-Post Results on Teacher Self-Efficacy Ratings

As seen in Figure 3, reported self-efficacy ratings of teachers in the Formal Observation Model dropped from pretest to posttest, whereas reported self-efficacy ratings of teachers in the Differentiated Supervision Model were relatively flat with the greatest increase seen in the Classroom Management factor.

Construct 3: Educator Effectiveness Ratings

With the passage of Act 82, Pennsylvania classroom teachers are to be given summative evaluations known as Educator Effectiveness ratings. Teachers are placed into one of two evaluation protocols, the Formal Observation Model or the Differentiated Supervision Model. Each teacher must participate in the Formal Observation Model once every cycle and in the Differentiated Supervision Model the remaining years of the cycle. School districts set the length of the cycle, suggested to be 3-5 years. Regardless of the protocol, teachers will be given ratings in each of the four domains of Danielson's *Framework for Teaching* rubric annually.

Act 82 proffers three examples of Differentiated Supervision modes for districts to consider: (1) the Peer Coaching Mode, in which teachers work together in pairs (or trios) to discuss their professional needs in the areas of pedagogy, student learning, and curriculum; (2) the Self-Directed Model/Action Research Mode, where teachers may work alone or in small groups to complete an action research project; or (3) the Portfolio Mode, where teachers examine their own practice and develop portfolios of artifacts and evidence documenting their level of competence in each domain of Danielson's rubric. In this research study, the Portfolio Mode was chosen to evaluate teachers in the Differentiated Supervision Model.

Summary data on educator effectiveness ratings. For the overall Educator Effectiveness ratings, 80% of classroom teachers were rated Proficient and 20% were rated Distinguished by their supervisor (see Table 27). No participants received an overall rating as Needs Improvement or Failing. However, there were a higher percentage of Distinguished ratings within the Differentiated Supervision participants (22%) than within the Formal Observation participants (14%).

Table 27

Overall Proficiency Ratings for Classroom Teachers

	Proficient	Distinguished
Formal Observation	30	5
Differentiated Supervision	59	17
Overall	89	22

When disaggregating proficiency levels by domains, similar Educator Effectiveness ratings occurred between the participants in the two evaluation protocols in Domain 1 (Planning and Preparation), as shown in Figure 4.

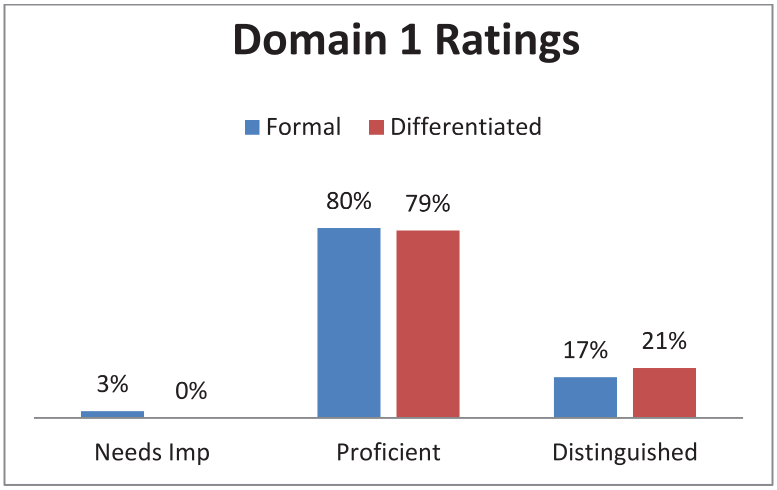


Figure 4. Comparison of Proficiency Ratings in Domain 1 by Evaluation Type

Domain 2 (The Classroom Environment) focuses on components observed during a taught lesson. The results of proficiency ratings in this domain, disaggregated by evaluation protocol, are shown in Figure 5.

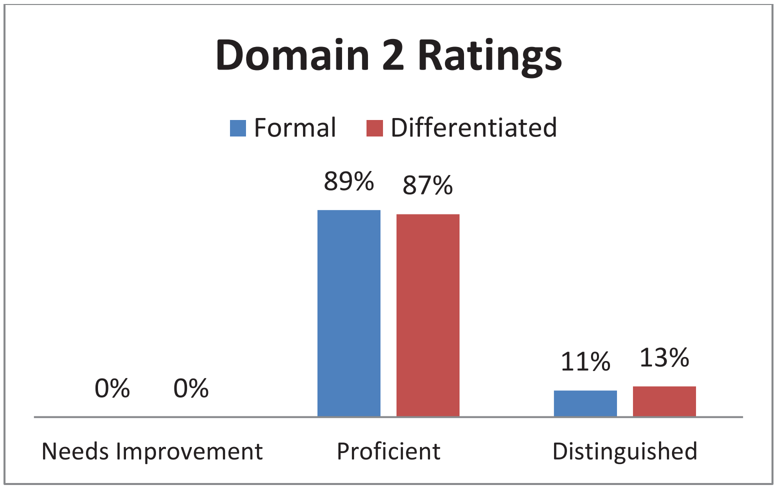


Figure 5. Comparison of Proficiency Ratings in Domain 2 by Evaluation Type

The components evaluated in Domain 3 (Instruction), refer to actual classroom instructional practices. A comparison of proficiency levels across the Domain 3 ratings are shown in Figure 6.

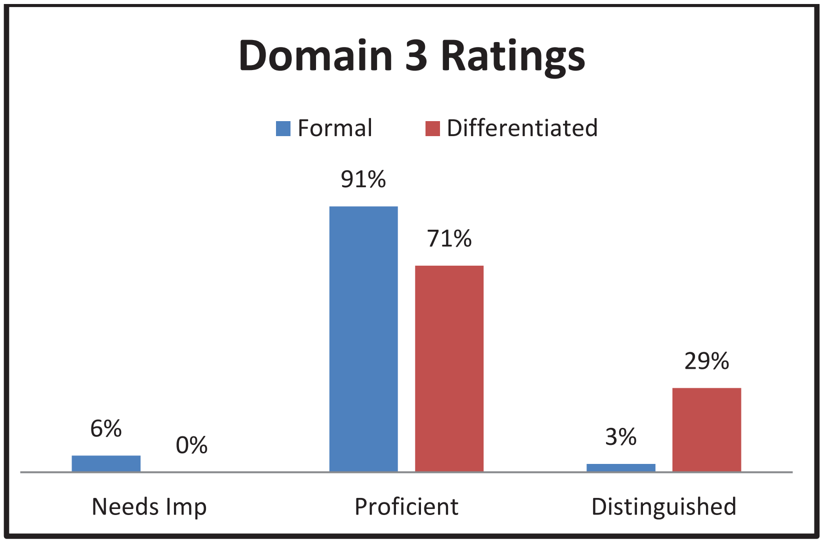


Figure 6. Comparison of Proficiency Level Ratings in Domain 3 by Evaluation Type

The components of Domain 4 (Professional Responsibilities) are not observable during classroom instruction. A comparison of the ratings between the two evaluation protocols is shown in Figure 7.

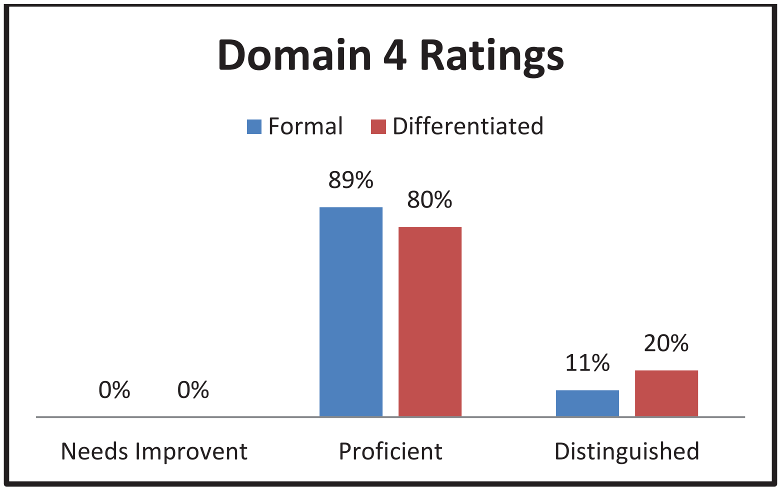


Figure 7. Comparison of Proficiency Ratings in Domain 4 by Evaluation Type

Descriptive analysis of educator effectiveness ratings. Descriptive analysis of the educator effectiveness ratings by domain was performed to assess the assumptions of normality. Summary statistics are found in Table 28 (output provided in Appendix V).

Table 28

Summary Statistics

Educator Effectiveness Ratings	N	M	SD	Skewness	Kurtosis
Domain 2					
Formal Observation	35	2.12	.24	1.18	1.32
Differentiated Supervision	76	2.23	.22	.69	-.32
Domain 3					
Formal Observation	35	2.03	.32	-.90	2.07
Differentiated Supervision	76	2.31	.28	.35	-1.00

Assumptions of normality. In a normal distribution, the values of skewness and kurtosis should approach zero. Since all variables have a skewness value less than |2.0|, the assumption of normality is tenable for all variables (Tabachnick & Fidell, 2013). The kurtosis values are all less than |7.0|, indicating the assumption of normality is tenable for all variables (Tabachnick & Fidell, 2013).

The Kolmogorov-Smirnov (K-S) tests of normality were significantly non-normal for ratings of participants in the Formal Observation Model for Domain 2, $D(35) = .26$, $p < .001$ and Domain 3, $D(35) = .18$, $p < .001$. For participants in the Differentiated Supervision Model, K-S tests of normality were also significantly non-normal: Domain 2, $D(76) = .19$, $p = .01$ and Domain 3, $D(76) = .21$, $p < .001$ (see output in Appendix W). As indicated above, the skewness and kurtosis of these variables are within the acceptable ranges; these significant 1 Sample K-S results are not concerning. The 1 Sample K-S test is sensitive to sample sizes in excess of $n = 100$, since it is based on a chi-square distribution (Tabachnick & Fidell, 2013). Therefore, normality of these variables is assumed tenable.

Homogeneity of variance. Levene's test of homogeneity of variance supports the assumption that the variances are not significantly different (see output in Appendix W). For Domain 2, $F(1,109) = .09$, $p = .77$ and Domain 3, $F(1,109) = .03$, $p = .86$.

Correlations. For Formal Observation Model participants, there was a slight, positive correlation between the Educator Effectiveness ratings participants received in Domains 2 and 3, and a strong, positive correlation between these Domains for participants in the Differentiated Supervision Model (see Table 29 and output in Appendix X).

Table 29

Correlations Comparing Educator Effectiveness Ratings in Domains 2 and 3

	r	p
Formal Observation Model Participants (*n* = 35)	.37*	.03
Differentiated Supervision Model Participants (*n* = 76)	.76**	< .001

** Correlation is significant at the .01 level (two-tailed)

* Correlation is significant at the .05 level (two-tailed)

Classroom Instructional Practices observed in round 2 were compared to Educator Effectiveness ratings generated by administrators. The only significant correlation occurred in Domain 3 with participants in the Differentiated Supervision Model (*r* = .35, *p* = .002).

Paired-samples t-tests. Ratings for Domain 2 (Classroom Environment) and Domain 3 (Instruction), were collected during Observation 2 and the Educator Effectiveness evaluations (see output in Appendix Y). As seen in Table 30, there were significant differences and a large effect size between the Observation 2 and Educator Effectiveness ratings of Domain 3 for Formal Observation participants, but no significant difference with their Domain 2 ratings. For Differentiated Supervision participants, there were significant differences in both Domains between their Observation 2 and Educator Effectiveness ratings.

Table 30

Paired-Sample t-Tests for Observation 2 and Educator Effectiveness Ratings

Variables	M	SE	*t*-statistic	r
Domain 2 Ratings				
Formal Observation				
Observation 2	2.13	.03	*t*(34) = .30, *p* = .77	.05
Educator Effectiveness	2.12	.04		
Differentiated Supervision				
Observation 2	2.15	.02	*t*(75) = -2.42, *p* = .02	.28
Educator Effectiveness	2.23	.03		
Domain 3 Ratings				
Formal Observation				
Observation 2	2.25	.03	*t*(34) = 3.79, *p* = .001	.54
Educator Effectiveness	2.03	.05		
Differentiated Supervision				
Observation 2	2.22	.03	*t*(75) = 2.58, *p* = .01	.29
Educator Effectiveness	2.31	.03		

Summary

Chapter four compares the Classroom Instructional Practices, Self-Efficacy ratings, and Educator Effectiveness scores of teachers participating in different evaluation protocols. To study the effect that different evaluation protocols may have on each of these constructs, various levels of data were used. Classroom Instructional Practices' data and Self-Efficacy data were used at the aggregate and pair level, whereas Educator Effectiveness data were used at the aggregate level.

Descriptive analysis of Classroom Instructional Practices reveals there were no significant differences between the ratings of participants in the two evaluation protocols during Observation 1. During Observation 2, there were no significant differences between the average ratings of Classroom Instructional Practices of the two groups, but when disaggregated by Domains, the Formal Observation Model participants had slightly lower ratings in Domain 2 and slightly higher ratings in Domain 3 than the Differentiated Supervision participants. In comparisons of Observation 1 to Observation 2 data, ratings of participants in the Formal Observation Model were not significantly different in either domain, whereas Domain 3 ratings of participants in the Differentiated Supervision Model were significantly greater in Observation 2 compared to Observation 1.

The within-subjects analysis of Classroom Instructional Practices indicates there is a significant increase in participants' scores between Observation 1 and Observation 2 for both Domain 2 and Domain 3. When examined by type of evaluation protocol, the two groups began with similar ratings in Domain 2 during Observation 1; however the Formal Observation participants show greater gains in this domain during Observation 2. Domain 3 ratings reveal a different trend: the two groups produced similar scores during Observation 2, but the Differentiated Supervision participants had much lower ratings in this Domain during Observation 1.

Descriptive analysis of Self-Efficacy data reveals Formal Observation participants report greater self-efficacy in all three subscales than participants in Differentiated Supervision during the pre-administration of the survey. The only significant difference between the two groups is found for the Student Engagement subscale. In the post-administration of the survey none of the subscale ratings had significant differences between the participants in the two evaluation protocols. The reported self-efficacy ratings of teachers in the Formal Observation Model dropped from pre-administration to post-administration of the survey, whereas reported self-efficacy ratings of teachers in the Differentiated Supervision Model were relatively flat with the greatest increase seen in the Classroom Management subscale.

Descriptive analysis of Educator Effectiveness data reveals a significant difference and a large effect size between the Observation 2 and Educator Effectiveness ratings of Domain 3 for Formal Observation participants, but no significant difference with their Domain 2 ratings. For Differentiated Supervision participants, there were significant differences in both Domains between their Observation 2 and Educator Effectiveness ratings.

Chapter 5

Discussion

Pennsylvania's Act 82 legislation mandated the implementation of a new teacher evaluation system, aligned to the Danielson framework. Beginning with the 2013-14 school year, the legislation requires school districts to eliminate the dichotomous teacher evaluation system (Satisfactory or Unsatisfactory) and implement a standards-based evaluation with all classroom teachers. Districts are required to implement a Formal Observation Model with a portion of their teachers and choose a mode of Differentiated Supervision for the rest of the teaching staff. Teachers are to be cycled through the Formal Observation Model over a period of years to be determined by the district. The Formal Observation Model is designed to include professional conversations between the teacher and supervisor through an elaborate process which includes a pre-conference, classroom observation, post-conference and follow-up walkthroughs. The Differentiated Supervision Model, used with teachers during the years they are not participating in the Formal Observation Model, must be aligned to the Danielson framework. Three options for this model were suggested, but not limited to: Peer Coaching, Self-Directed/Action Research, or Portfolio modes (PDE, 2013). The administrators in the school district in this study selected the Portfolio Mode, in which teachers examine their own practice and share artifacts or evidence of their performance level in each domain of the Danielson framework (PDE, 2013).

Regardless of the evaluation model, teachers receive a rating for each of the four domains of the Danielson framework: (1) Planning and Preparation, (2) Classroom Environment, (3) Instruction, and (4) Professionalism. The new evaluation system will generate an Educator Effectiveness rating, based on a zero-to-three point rubric representing the teacher's overall level of proficiency aligned to Charlotte Danielson's *Framework for Teaching*: (0) Failing, (1) Needs Improvement, (2) Proficient, or (3) Distinguished.

The purpose of this study was to explore the impact of the new PA state-mandated, high-stakes teacher evaluation model on the use of classroom instructional practices by teacher participants. The Commonwealth's option for school districts to introduce two models of the evaluation protocol provided a unique opportunity to make comparisons of the classroom practices of the participants in these models. Since this study was focused on teachers' classroom instructional practices, only ratings in the domains where practices can be directly observed (Domain 2 and Domain 3) were examined. The research investigated the potential of the two evaluation models to impact instructional practices in classroom instruction. This study examined the relationships among measurements of teachers' classroom instructional practices and beliefs, and sought to answer the following questions:

1. Did the Classroom Instructional Practices of teachers in the Formal Observation Model improve to a greater extent than teachers in the Differentiated Supervision Model over the course of the year?
2. Was there a relationship between teachers' Self-Efficacy scores and their participation in one of the two evaluation models?
3. How did the summative Educator Effectiveness ratings of teachers in each evaluation protocol compare? Was there a relationship between the

summative Educator Effectiveness ratings of teachers and the observed ratings of their Classroom Instructional Practices?

Construct 1: Classroom Instructional Practices

The first area explored in this study was the potential for the new state-mandated evaluation model in PA to have a positive impact on a relevant construct, teachers' Classroom Instructional Practices. Thirty-five teachers were placed in the Formal Observation Model and 76 teachers in the Differentiated Supervision Model. The dependent variable, Classroom Instructional Practices, was rated by the researcher during unannounced observations of the teacher's instruction. The observations were conducted twice, during the first- and fourth-nine weeks of the school year.

Findings. Descriptive analysis of Classroom Instructional Practices revealed the ratings in both groups improved between Observation 1 and Observation 2. Using a paired-samples *t*-test, the average increase in the ratings of Formal Observation participants (.05) was not significant; the average increase in the ratings of Differentiated Supervision participants (.08) was significant (p = .016).

Finer-grain comparisons were made by disaggregating observation ratings by domains. Domain 2 consists of components and elements in the area of Classroom Environment and Domain 3 refers to Instruction. The results indicate:

1. For Formal Observation Model participants, there was essentially no change in Domain 2 (Classroom Environment) ratings; ratings in Domain 3 (Instruction) did improve (2.16 vs. 2.25), although the difference was not significant.
2. Ratings of Differentiated Supervision Model participants improved in both domains; however, only changes in Domain 3 scores were significant (p = .016).

The within-subjects analysis of Classroom Instructional Practices indicates there was a significant increase in aggregate teachers' scores between Observation 1 and Observation 2 for both Domain 2 and Domain 3. When examined by type of evaluation protocol, the two groups began with similar ratings in Domain 2 during Observation 1; however, the Formal Observation participants showed greater gains in this domain during Observation 2. Domain 3 ratings revealed a different trend: the two groups produced similar scores during Observation 2, but the Differentiated Supervision participants had much lower ratings in this Domain during Observation 1.

Implications. The heart of this study was to determine if there was a significant difference in use of best-practices in classroom instruction as a result of the implementation of two different evaluation models. Specifically, the following question was posed:

Did the Classroom Instructional Practices of teachers in the Formal Observation Model improve to a greater extent than teachers in the Differentiated Supervision Model over the course of the year?

The primary purpose for teacher evaluation is to ensure that expectations of the public for high-quality teachers in their schools are met. However, teacher evaluation can serve another purpose: the promotion of professional learning (Danielson, 2012). This goal is attainable if purposeful, professional conversations between teachers and their supervisors occur in conjunction with formal or informal observations (Danielson, 2012).

As implemented in the district of study, only the Formal Observation Model embedded the professional conversations between teacher and administrator through the course of the school year. The teachers in the Differentiated Supervision Model engaged in examinations of their practice in order to gather the required documentation and evidence for their portfolios, but they did not participate in formal, collegial discussions of their practice until the actual portfolio presentations. It was hypothesized that teachers participating in the Formal Observation Model would implement changes and improvements in classroom instructional practices to a greater extent than teachers participating in the Differentiated Supervision Model. However, when comparing the overall ratings for the initial and final observations, the null hypothesis could not be rejected as there was no significant difference in the improvement of instructional practices with the Formal Observation participants. Unexpectedly, the improvement in ratings of the Differentiated Supervision participants was significant, in both the overall observation and Domain 3 ratings. To understand the unanticipated findings, a review of portfolio research provides further insight.

Danielson (2012) contended that a teacher evaluation system can promote professional learning by embedding activities such as self-assessment and reflection on practice. A portfolio can serve as a self-assessment, providing teachers the opportunity for focused reflection on their teaching practices (Berrill & Whalen, 2007; Riggs & Sandlin, 2000). Even the use of portfolios for summative evaluation fosters self-reflection about possible changes in practice (Knapper & Wright, 2001). In the process of selecting evidence aligned to a standards-based portfolio, teachers may be alerted to areas of weakness if they are unable to find evidence to the contrary (Gelfer et al., 2004). Teachers acknowledged that the portfolio process “encouraged them to clarify their instructional goals and more closely examine their teaching practices” (Wolf et al., 1996, p. 285). Despite these positive sentiments regarding the value of portfolios to impact professional growth, in a survey of 600 teachers involved in a three-year pilot implementation of portfolios, teachers reported improvement in “self-reflection as a result of portfolios but the self-reflection had little impact on teaching practice” (Tucker et al., 2003, p. 591). The significant improvement in the observed classroom instructional practices (Domain 3) of teachers in the portfolio mode of this study stands in direct contrast to the findings of Tucker et al. (2003). Several reasons may account for this disparity: (1) while teachers in the Tucker et al. study expressed the opinion that self-reflection had no impact on teaching practice, actual observations of their teaching practices were not conducted to confirm these opinions, and (2) the use of portfolios in the Tucker et al. pilot study was not attached to a high-stakes summative evaluation for the participants. The use of these two evaluation models for summative purposes may explain some of the differences in the expected results.

Limitations. Due to the constraints imposed by Act 82 legislation, namely, that all non-tenured teachers should be placed into the Formal Observation Model, completely random assignment of teachers into each evaluation model was not possible. While there was no significant difference in the initial ratings of classroom instructional practices of the two groups of teachers in different evaluation protocols, there could be factors related to a teacher’s non-tenured status that affected the final ratings of teachers in the Formal Observation group.

Additional data analysis was conducted to examine differences in the ratings of classroom instructional practices of tenured teachers (*n* = 22) and non-tenured teachers (*n* = 13) in the Formal Observation Model. As shown in Appendix Z, an independent samples *t*-test reveals no significant difference between the ratings of these two groups of teachers during Observation 1, *t*(33) = .54, *p* = .60, or Observation 2, *t*(33) = .86, *p* = .39.

To compare changes in ratings of tenured teachers' Classroom Instructional Practices in the two evaluation protocols, a paired-samples *t*-test was conducted. There were 22 tenured teachers in the Formal Observation Model and 76 tenured teachers in the Differentiated Supervision Model. As shown in Appendix Z, the classroom practices of all tenured teachers improved in each Domain between Observation 1 and Observation 2. However, the improvements for tenured teachers in the Formal Observation Model were not significant: for Domain 2, *t*(21) = .48, *p* = .64; for Domain 3, *t*(21) = 1.5, *p* = .15; and for changes in Overall Observation ratings, *t*(21) = 1.04, *p* = .31. As reported earlier, there were significant improvements in the ratings of Classroom Instructional Practices of teachers in the Differentiated Supervision model in Domain 3, *t*(75) = 2.46, *p* = .016 and in their Overall Observation ratings, *t*(75) = 2.46, *p* = .016. The lack of statistical significance in the ratings of tenured teachers in the Formal Observation Model could be a result of the reduced sample size.

Reliability of measures of classroom instructional practices was limited by the use of only one research observer in this study. Unfortunately, the addition of a second observer for the amount of time necessary to conduct 222 observations for full-periods of instruction was not possible. All other administrators were assigned to conduct the individual evaluations of teachers in the two groups.

Recommendations for practice. Recent research on the link between high quality teacher professional development and resulting improvements both in teaching skills and student achievement are relevant and important to the findings in this study.

Antoniou and Kyriakides (2013) found that teachers participating in a professional development model that focused critical reflections on their individual areas of need (depending on their present skill level and experience) improved their teaching skills more than teachers who experienced a holistic approach, in which teachers reflected on "any aspect of their teaching practice irrespective of the stage at which they were situated" (p. 9). These results are important considerations for PA district leaders who must choose among various options for the Differentiated Supervision evaluation of teachers. If the overarching purpose for evaluation is to improve instructional practices, selecting a model that incorporates important components of professional development is beneficial to all stakeholders.

Professional growth of the teacher is maximized if professional dialogue occurs about the contents of the portfolio and the standards they represent (Gelfer, Xu, & Perkins, 2004; Riggs & Sandlin, 2000). In this first year of implementation, school administrators did not provide specific guidance or directives to teachers regarding the collection of evidence for portfolios. Since the participants in the Portfolio Mode had significantly improved ratings in their Classroom Instructional Practices in Domain 3, an additional focus on the standards embedded in Domain 2 may result in improved practices in this domain as well.

Researchers caution against the use of an evaluation process for both formative and summative purposes. Marzano (2012b) posited that inherently different systems are needed for an evaluation system focused on developing teachers and improving learning than a system focused on measuring teacher competence. Difficulties arise "in integrating the requirements of an evaluation policy geared toward job status decisions with those of a policy aimed at improving teaching" (Darling-Hammond, Wise, & Pease, 1983, p. 287). The high-stakes attributable to the evaluations of teachers in the two models in this study could be a major factor in the study's findings. The Formal Observation participants were highly focused on satisfactory performance for one intense observation period, with no additional requirements to examine their daily practices throughout the year. Perhaps the inclusion of some of the components of the Portfolio Mode for these participants will enhance classroom practices. Collection of evidence and artifacts relevant to the Danielson framework, separate from the discussions embedded in the Formal Observation process, can be incorporated into this cycle of evaluation.

Future research. According to Kimball (2002), an evaluation model that includes a significant amount of discourse between the teacher and the evaluator is likely to enhance teacher reflection and growth. In this research study, the classroom practices of teachers in the Portfolio Mode significantly improved, while the improvement in practices of teachers in the Formal Observation Model was not significant. The unexpected results lead to several potential questions or areas for further research:

1. Is there a relationship between the constructs of classroom instructional practices and teacher reflection and growth?
2. Can professional discourse between colleagues on classroom practices be as effective for teacher growth as discourse between teacher and evaluator?
3. Does the professional experience (i.e. years of service, grade level/subject area assignment) of the educator have an impact on the findings of this study?
4. How do the classroom instructional practices of teachers evaluated in another mode of Differentiated Supervision (e.g., Peer Coaching, Action Research)

compare to those of teachers in the Formal Observation Model or the Portfolio Mode of Differentiated Supervision?

Construct 2: Teacher Self-Efficacy Ratings

Teacher self-efficacy, defined by Klassen, Tze, Betts, & Gordon (2011) as the "confidence teachers hold about their individual and collective capability to influence student learning" (p. 21), is believed to influence teachers' professional behaviors (Holzberger, Philipp, & Kunter, 2013). Teachers' perceptions of their own abilities and the contexts in which they teach both influence and are influenced by their environment (Fives & Buehl, 2010; Tschannen-Moran & Hoy, 2001). The teachers in this research study were essentially in the same environment, although in two separate schools (one elementary and one high school). Approximately equal numbers of teachers in each building were placed into each of the two evaluation protocols. This study examined the possible differences in teachers' perceptions of their abilities based on their experience with different evaluation protocols.

Findings. A greater percentage of teachers in the Formal Observation Model responded to both distributions of the survey: 89% Formal (*n* = 31) vs. 75% Differentiated Supervision (*n* = 53) participants in the first survey; in the second survey, 70% Formal (*n* = 26) vs. 64% Differentiated Supervision (*n* = 49) participants responded. The survey measures three aspects of teacher reported Self-Efficacy: Student Engagement, Instructional Strategies, and Classroom Management.

Descriptive analysis of Self-Efficacy data revealed Formal Observation participants reported greater self-efficacy in all three subscales than participants in Differentiated Supervision during the pre-administration of the survey; however, only the difference in the Student Engagement subscale was significant. In the post-administration of the survey, none of the subscale ratings had significant differences between the participants in the two evaluation protocols. The reported self-efficacy ratings of teachers in the Formal Observation Model dropped from the pre-administration to the post-administration of the survey, whereas reported self-efficacy ratings of teachers in the Differentiated Supervision Model were relatively flat with the greatest increase seen in the Classroom Management subscale.

Implications. Based on prior research that found teachers' self-efficacy ratings from the beginning of the school year were significantly related to the teachers' classroom performance ratings measured at the end of the year (Heneman, Kimball, & Milanowski, 2006), this study sought to explore the question:

Was there a relationship between teachers' Self-Efficacy scores and their participation in one of the two evaluation models?

In the current research study, there was a significant difference between the composite self-efficacy ratings (an average of the three subscales) of the participants in the two evaluation protocols measured at the beginning of the school year. Formal Observation Model participants' composite self-efficacy ratings (7.54) were significantly higher than Differentiated Supervision participants' ratings (7.16).

If the Heneman et al. (2006) research results are applicable, there should be a significant difference between the end-of-year performance ratings of the two evaluation protocol groups. However, there was no significant difference between the final performance ratings (each group had an average performance rating of 2.19 based on measures of Classroom Instructional Practices by the researcher in the final round of observations). Furthermore, there were no significant differences in the final performance ratings of the two groups in either Domain 2, $t(109) = -.45$, $p = .65$, or Domain 3, $t(109) = .62$, $p = .54$.

Based on the finding that teachers in the two evaluation protocols had significantly different ratings on one subscale of the initial Self-Efficacy survey, further analysis was conducted to determine if there was any difference in a related performance measure. Teachers in the Formal Observation Model rated themselves significantly higher in Student Engagement than the Differentiated Supervision participants (7.17 vs. 6.58). In the Danielson framework, one component (3C - Engaging Students in Learning) was selected to determine if Formal Observation participants had higher performance ratings than Differentiated Supervision participants. There was no significant difference (Formal Observation, $M = 2.23$; Differentiated Supervision, $M = 2.22$, $p = .94$, as shown in Appendix Z).

While these research results appear to conflict with the Heneman et al. (2006) findings, there were distinct differences in the methodology used in the two studies which could account for the disparities. Heneman et al. (2006) measured teachers' performances by using components from Domains 1 and 3 of the Danielson *Framework for Teaching*, whereas, the present research study used the components of Domains 2 and 3. The components of Domain 1 (Planning and Preparation) were not observable during a classroom observation and therefore were not included in the measure of teachers' performances in this study.

Another difference occurred with the relative sampling processes of the two studies. Heneman et al. (2006) used the self-efficacy surveys from 1,075 teachers across all grade levels and subjects (p. 7), but used the performance measures of 180 elementary teachers (p. 8). In the present study, performance measures were obtained for all 111 classroom teachers, but 75% of these teachers (*n* = 84) were collected. Heneman et al.'s (2006) argument for the existence of a significant relationship between self-efficacy ratings and teacher performance measures might be stronger if there was a closer match between the two sample populations used in the study (i.e. restricting the self-efficacy survey results to elementary teachers).

An early study on the relationship between teachers' sense of efficacy and their perspectives about student behavior and classroom management provided the following results: teachers with low efficacy scores chose classroom management strategies such as punishment and coercion to control student behavior, whereas teachers with high efficacy scores reported less occurrence of student disruptions and more positive ways of handling misbehavior (Woolfolk, Rosoff, & Hoy, 1990). There may be a reciprocal relationship at work, where teachers' levels of self-efficacy impact, and in turn, are impacted by student behavior. In a recent study, Holzberger et al. (2013) found that teachers, regardless of years of experience, modified their own self-efficacy beliefs over the course of a school year, and increases in self-efficacy ratings occurred in response to "experiences of success in the classroom" (p. 783). Again, there seems to be a reciprocal effect between the two constructs (Holzberger et al., 2013). Notably, the Holzberger et al. (2013) study used a self-efficacy instrument and performance measure unrelated to the tools used in this study.

In light of the current study's findings that self-efficacy ratings of Formal Observation participants decreased in all three subscales from pre-administration to post-administration, perhaps their experiences in the Formal Observation process contributed to these changes. An instrument that rates the participants' experiences is necessary to conduct such analyses. This was beyond the scope of this study.

Limitations. The anonymous and voluntary implementation of the Self-Efficacy survey restricted the use of the results for making statistically significant comparisons between the participants of the two evaluation models. Anonymous surveys are more likely to generate honest responses from the volunteers, but prevent analyses that link the classroom performances of individual participants to their self-efficacy ratings. Furthermore, biased results may be introduced if non-volunteers are "different from the rest of the population in ways that affect the survey answers" (Fowler, 2014, p. 10).

Therefore, the lack of significant changes in self-efficacy results of teachers in either evaluation protocol (as measured at the beginning and end of the school year) may have been a result of the bias introduced by the voluntary nature of the survey. This may also have caused the initial significant differences between the self-efficacy ratings of the participants in the two evaluation models.

Recommendations for practice. Teachers' perceptions of their own abilities and the contexts in which they teach both influence and are influenced by their environment, which in turn are thought to affect their classroom behaviors (Fives & Buehl, 2010; Tschannen-Moran & Hoy, 2001). The potential existence of a reciprocal effect between teachers' self-efficacy beliefs and the evaluation model they experience, coupled with the perceived positive relationship between self-efficacy and classroom practices, is an important consideration for practicing educators. If self-efficacy beliefs are impacted by the evaluation process itself, ways to mitigate any negative influence should be explored.

Future research. According to Heneman et al. (2006), higher self-efficacy ratings at the beginning of the year should reflect higher end-of-year performance ratings. The lack of significant differences in performance ratings by Formal Observation participants in this study, despite their higher, initial self-efficacy scores, provides a basis for further investigation. The potential for bias in self-efficacy ratings generated in voluntary surveys is an important consideration in proposed studies.

In prior research of teacher self-efficacy ratings, Heneman et al. (2006) measured teacher performance in components of Domains 1 and 3, while this research study used components of Domains 2 and 3. Hence, a review of the alignment of components in either domain to the three sub-scales of the Self-Efficacy survey may reveal significant relationships. Domain 4, Professional Responsibilities, makes up a significant portion of the Danielson framework. Studies exploring the relationship between the components of this domain and teacher self-efficacy ratings may be informative.

Based on prior research, teachers' perceptions of their own abilities and the contexts in which they teach both influence and are influenced by their environment, which in turn are thought to affect their classroom behaviors (Fives & Buehl, 2010; Tschannen-Moran & Hoy, 2001). In the current study, Formal Observation participants reported a decrease in self-efficacy scores across the school year. Researchers may want to determine if the experiences of participants in the Formal Observation Model contributed to these changes in their self-efficacy ratings.

Construct 3: Educator Effectiveness Ratings

The third area investigated in this study was the relationship between teachers' Classroom Instructional Practices and their Educator Effectiveness ratings. Teachers in the Formal Observation Model received a rating based on the clinical observation process, conducted between the teacher and an administrative supervisor, and included a pre-conference, scheduled classroom observation, and a post-conference. Teachers in the Differentiated Supervision Model collected artifacts and evidence across the four domains of the Danielson framework in preparation for a final presentation to administrative supervisors as a portfolio. Six administrators conducted the evaluations with different groups of teachers. Since Educator Effectiveness ratings were only completed once (at the end of the school year), only Classroom Instructional Practices' ratings collected in Observation 2 were used in comparative analyses.

Findings. Descriptive analysis of the data revealed a significant difference and a large effect size between Observation 2 ratings and Educator Effectiveness ratings of Domain 3 for Formal Observation participants, but no significant difference with their Domain 2 ratings. For Differentiated Supervision participants, there were significant differences in both Domains between their Observation 2 and Educator Effectiveness ratings.

Principals/supervisors rated Formal Observation teachers significantly lower than the researcher in Domain 3 (2.03 vs. 2.25), but rated Differentiation Supervision teachers significantly higher in Domain 2 (2.23 vs. 2.15) and Domain 3 (2.31 vs. 2.22).

Implications. Educator Effectiveness ratings were determined through two different protocols (the pre-observation conference, observation, and post-observation conference for Formal Observation participants, and the Portfolio presentations for Differentiated Supervision participants). A comparison of the proficiency ratings obtained by participants in these two protocols is of interest to school administrators and teachers alike. Therefore, the following question was explored:

How did the Educator Effectiveness ratings of teachers in the Formal Observation Model compare to the ratings of teachers in the Differentiated Supervision Model?

According to Act 81 legislation, the four domains of the evaluation protocol are weighted: 20% for Domain 1, 30% for Domain 2, 30% for Domain 3, and 20% for Domain 4. For the overall Educator Effectiveness ratings, 80% of classroom teachers were rated Proficient and 20% were rated Distinguished by their supervisor. No participants received an overall rating as Needs Improvement or Failing. However, there were a higher percentage of Distinguished ratings within the Differentiated Supervision participants (22%) than within the Formal Observation participants (14%).

One of the goals of the new teacher evaluation process in PA is to create a system that makes finer distinctions than the decades-long practice of rating teachers with a dichotomous satisfactory/unsatisfactory model. Although 100% of the teachers in this district received a rating of Proficient or higher evaluation in this first year of implementation of the new evaluation system, these results represent 85% of the teacher's final Educator Effectiveness scores. The other 15% will not be added until early September. This additional component, the building-level School Performance Profile score, consists of a complex set of formulae representing extensive calculations of various student academic achievement factors.

Administrators consistently rated the teachers in the Differentiated Supervision Model higher than teachers in the Formal Observation Model. Although it could be posited that the cohort of teachers in the Formal Observation Model did not perform as well in Domains 2 and 3 than teachers in the Differentiated Supervision Model, a more likely explanation lies in the differences in the processes producing these ratings. Formal Observation ratings were collected during full-period classroom visits, while Differentiated Supervision ratings were evaluated during the brief portfolio presentations.

When disaggregating proficiency levels by domains, similar Educator Effectiveness ratings occur between the participants in the two evaluation protocols in Domain 1 (Planning and Preparation). A lower percentage of Formal Observation Model participants were rated Distinguished when compared to participants in the Differentiated Supervision Model (17% vs. 21%). None of the Differentiated Supervision teachers were rated Needs Improvement, but a small percentage (3%) of Formal Observation teachers received this rating in Domain 1.

Teachers in the Differentiated Supervision Model were evaluated on their presentation of artifacts or evidence that demonstrated their skills or knowledge in meeting these components. This was done in a group setting, with a limited amount of time for each teacher to discuss the evidence. However, teachers in the Formal Observation Model met one-on-one with their supervisor for a full period (42 minutes) during the Pre-Conference of this protocol. With the significant difference in time devoted to Formal Observation participants, a more reliable and valid assessment of their ratings in this domain is probable.

Domain 2 (The Classroom Environment) focuses on components observed during a taught lesson. There is little difference between the supervisors' direct observations of these components in the classrooms of teachers (the Formal Observation Model participants) and the supervisors' evaluation of artifacts and evidence presented by the Differentiated Supervision participants.

The components evaluated in Domain 3 (Instruction), referred to actual classroom instructional practices. In a comparison of proficiency levels across the Domain 3 ratings, a much higher percentage of Differentiated Supervision participants were scored as Distinguished (29%) compared to the number of Formal Observation participants (2.9%). In addition, no teachers in the Differentiated Supervision Model were rated below Proficient in this Domain, but almost 6% of Formal Observation participants were rated as Needs Improvement.

This domain was at the heart of the Formal Observation protocol. All the planning and preparation done in Domain 1 was linked directly to the observable elements of Domain 3. While Differentiated Supervision participants can select their own evidence for evaluation, Formal Observation participants must demonstrate these competencies during a live observation. For example, teachers in the Formal Observation Model may have described methods for differentiating instruction for students when they discussed their planning during the pre-observation conference, but unless the supervisor observes the differentiation during the lesson, the teacher's rating will be lowered. This type of comparison between planning and implementation of classroom instructional practices is not possible for Differentiated Supervision evaluations. Therefore, it is not surprising that the Formal Observation cohort had some Needs Improvement ratings in this domain (which are unsurprisingly similar to Domain 1 ratings).

The components of Domain 4 (Professional Responsibilities) were not observable during classroom instruction. For teachers in the Formal Observation Model, these components were rated during the one-on-one post-observation conferences with their supervisors. All participants in each evaluation protocol were rated Proficient or higher. A lower percentage of Formal Observation participants received a Distinguished rating compared to the percentage of participants in the Differentiated Supervision Model (11% vs. 20%).

This study also investigated a possible relationship between the Educator Effectiveness ratings completed by supervisors and observed ratings of their Classroom Instructional Practices (labeled Observation 2) completed in this research study. The relevant question was:

How did the Educator Effectiveness ratings of teachers in each evaluation protocol compare to the ratings of their Classroom Instructional Practices?

The Educator Effectiveness ratings for the 35 teachers in the Formal Observation Model were generated by principals conducting a classroom observation, as prescribed in the evaluation protocol. Observation 2 ratings were conducted by the researcher during the final round of classroom visits. There were two findings from the comparisons for Formal Observation participants:

1. There was no significant difference between ratings in Domain 2, Classroom Environment. The researcher's ratings had a mean of 2.13, and the combined mean of the six administrators was 2.12.
2. There was a significant difference (p = .001), and a large effect size (r = .54) between the mean Observation 2 and Educator Effectiveness ratings for the Formal Observation participants in Domain 3, Instruction (Obs. 2, M = 2.25; EE, M = 2.03).

The administrators conducted their respective Formal Observations during scheduled visits, whereas the researcher's observations were unannounced. If the unannounced observations are more likely to be representative of the teacher's actual classroom practices, the higher ratings obtained by the researcher were somewhat unexpected. There are several plausible explanations for this discrepancy.

Perhaps the tense and stressful nature of a formal observation impacts a teacher's instructional practices. Since 37% of these teachers are non-tenured, the formal observation is an even more formidable high-stakes' experience. However, if such an impact was in effect, why did it not occur with Domain 2 ratings? Testing this assumption by disaggregating the Formal Observation group according to tenure was not feasible, as sample size would be seriously compromised.

Another plausible explanation might be to assume the researcher's ratings are simply skewed towards a more positive direction; however, such a trend did not appear in Domain 2 ratings or across the comparisons with teachers in the Differentiated Supervision Model. Disaggregating ratings by individual administrator might show skewness attributable to one of these evaluators, but again, this was not a feasible option as sample sizes would be reduce dramatically.

Educator Effectiveness ratings for the 76 teachers in the Differentiated Supervision Model were generated during the portfolio presentations of these teachers, conducted in group settings of 18 - 20 colleagues. Observation 2 ratings of these teachers were collected by the researcher during the final round of unannounced classroom visits. There were two notable findings:

1. For Domain 2 (Classroom Environment), there was a significant difference (*p* = .02) between the researcher's classroom observations (*M* = 2.15) and the administrators' portfolio ratings (*M* = 2.23).
2. In Domain 3 (Instruction), there was a significant difference (*p* = .01) between the researcher's classroom observations of participants (*M* = 2.22) and the administrators' portfolio ratings (*M* = 2.31).

The fact that all the evidence for the administrators' ratings was presented by teachers and not observed directly by the administrators would likely result in higher ratings. This was confirmed with the significantly higher ratings given by administrators during the brief portfolio presentations. The researcher's ratings were based on direct, full-period observations of each teacher, which could result in more valid evaluations.

Limitations. The major weaknesses of this research protocol include various threats to validity and reliability, generalizability, and sample size. Social threats to internal reliability were possible since all participants were part of the same faculty. Generalizability of findings to other populations was limited by differences in various demographical and contextual factors of other populations. The power and effect size of the findings could have been diminished by the small sample sizes in this study. Additional limitations, specific to the instruments used in this study, are addressed below.

Educator Effectiveness ratings. Inter-rater reliability threats may have occurred with Educator Effectiveness ratings, as they were collected by different administrators and used different instruments for the two evaluation protocols. Kimball and Milanowski (2009) studied differences in evaluator decision-making to determine plausible explanations for differential validity across principals. If evaluators are intent on maintaining good, working relationships with teachers, they may be hesitant to provide negative feedback (Kimball & Milanowski, 2009).

Portfolios. Although administrators have been trained and tested on their ability to discern among proficiency levels within the Danielson framework (as applied to teachers in the Formal Observation Model), teachers in the Differentiated Supervision Model were evaluated with a Portfolio rubric created by the administrators. As a result, the instrument may have lacked construct validity. Portfolios that represent a comprehensive picture of teaching are believed to have face validity (Knapper & Wright, 2001); however, the comprehensiveness of the portfolio used in this study has not been evaluated.

Concurrent validity of an instrument may exist if evaluators are not able to distinguish between the four levels of proficiency. Although Tucker et al. (2003) found much greater differentiation occurred when the final ratings produced with portfolio evaluations were compared to prior evaluations based on traditional observations alone, the opposite results were obtained in this study (there was greater differentiation in the Educator Effectiveness ratings for teachers evaluated with a formal observation).

Consistency and subjectivity in portfolio ratings are important factors. Reliability of portfolio assessments may be affected by “subjective impressions and personal relationships between the rater and the teacher assessed” (Van der Schaff, Stolling, & Verloop, 2005, p. 47). Reliability can be enhanced if evaluators are trained on the types of evidence that are relevant to the purpose of the portfolio and if specific criteria regarding this evidence are developed (Johnson et al., 2000). Concerns with accuracy can be addressed by administrators conducting regular classroom observations, looking for evidence to support portfolio presentations (Attinello et al., 2006).

Recommendations for practice. There are several components of the Educator Effectiveness construct that are important to the overall process. The recommendations for improvement in each component are described below.

Classroom observations. A single observation score is dependent on various classroom factors that may not be indicative of the teacher's actual effectiveness. Hence, single observations are likely to produce inaccurate indicators of a teacher's classroom practice. Instead, averaging scores over multiple observations will improve the reliability of the evaluation.

Announced classroom observations can prevent the observer from viewing the day-to-day classroom experiences of students. As a result, evaluations are not honest reflections of classroom interactions, "and are not helpful for improving mediocre and ineffective teaching practices" (Marshall, 2012, p. 50). The obvious solution is to schedule multiple unannounced visits to capture the most accurate representation of teacher effectiveness.

One recommendation to address the inter-rater reliability of the observers is the implementation of training and certification of observers. Reliable evaluations of a teacher's practice should include multiple observations in order to capture an accurate picture of the large number of competencies and skills required of effective teachers. Reliability should be monitored by incorporating periodic observations by multiple, impartial observers.

Portfolios. Unlike the brief snapshots available during a single observation, portfolios provide administrators the opportunity to look closely at a practice as it unfolds over time and encourages the "reflection on those variables not easily captured during classroom observation" (Riggs & Sandlin, 2000, p. 24). It was suggested that administrators are better able to recognize differences in teacher performance with the additional insight into instruction provided by portfolios (Tucker et al., 2003). If artifacts are accompanied with explanations on their relationship to teaching, administrators gain deeper insight into the teacher's practices and instructional philosophies (Wolf et al., 1996; Tucker et al., 2003).

While portfolios may highlight excellence in teaching practices, the lack of uniformity in the portfolio structure makes it difficult to make fair comparisons (Peterson et al., 2001). To be relevant, portfolio evaluation must be based on specific criteria and aligned with particular standards and important classroom practices (Riggs & Sandlin, 2000) to prevent a miscellaneous collection of artifacts that have no "relationship to critical thinking or teacher reflection" (Blake et al., 1995, p. 44). Providing teachers with a model of an exemplar portfolio can assist them with selection of artifacts and evidence representing key concepts (Moore & Bond, 2002). Administrators can support teachers by providing them ongoing feedback during the process (Moore & Bond, 2002) and adequate time to develop and reflect on portfolio contents (Attinello et al., 2006; Tucker et al., 2003).

Walkthroughs. While the legislation permits two models of evaluation (Formal Observation or Differentiated Supervision), both aligned to the Danielson framework, regular classroom walkthroughs are suggested to support administrators' summative ratings. Unannounced walkthroughs are recommended after the Formal Observation Model to verify the teacher's implementation of suggested improvements in classroom practices that arose during the post-observation conference. Walkthroughs are also recommended for all teachers in either evaluation model throughout the school year for both formative and summative purposes.

Walkthroughs provide principals the opportunity to develop professional learning communities, and work collaboratively with staff to reflect and analyze their own instructional practices (Cotton, 2003; Downey et al., 2004; Kachur et al., 2010; Stronge et al., 2008). When conducted on a regular basis, classroom walkthrough data can be used to illuminate how teachers make curricular and instructional decisions (Downey et al, 2004). Ultimately, regular walkthroughs generate information to help teachers analyze their teaching practices (ASCD, 2007) and improve student achievement (Kachur et al., 2010).

Future research. Little research in the use of portfolios for professional development of practicing teachers (Berrill & Whalen, 2007) and teacher evaluation (Xu, 2003) has been conducted. The significant improvement in the Classroom Instructional Practices of teachers in the Portfolio Mode provides a strong basis for further investigation, particularly in its usefulness as an evaluation tool. The decreased differentiation in teachers' ratings in the Portfolio Mode is an important area for future research, since the impetus behind the new state-mandated evaluation process is to distinguish educators across four levels of proficiency.

Research regarding assessment of the contents of portfolios has been scant, and limited information on the reliability and validity of evaluators' ratings exists (Centra, 2000; Tucker et al., 2003). If improvements in teaching practices are the ultimate goal of an evaluation process, it is imperative that portfolio ratings measure the relevant components of teaching and learning. Just as important, if portfolio ratings are used to make high-stakes' summative decisions regarding a teacher's professional status, the reliability of the evaluators' ratings are critical.

Although not part of this study, Pennsylvania's newly adopted evaluation system permits the use of other modes besides portfolios for Differentiated Supervision. The Peer Coaching Mode, Self-Directed/Action Research Mode, and an alternative approved in advance for use by districts may be implemented during the years in which teachers are not participating in the Formal Observation Model. Research regarding the validity and reliability of these modes, and their potential impact on teaching and learning, is a significant concern for all stakeholders.

Summary

Over the last decade, policymakers and educational reform leaders have been investigating the potential of teacher evaluation models to improve student achievement. In a review of related literatures, Hallinger, Heck, and Murphy (2014) observed that the "'policy logic' driving teacher evaluation remains considerably stronger than empirical evidence of positive results" (p. 21). Insufficient evidence exists to support the premise that the latest generation of teacher evaluation systems is associated with "capacity development of teachers or more consistent growth in the learning outcomes of students" (Hallinger et al., 2014, p. 22).

The purpose of this study was to explore the impact of the new Pennsylvania state-mandated, high-stakes teacher evaluation model on the use of classroom instructional practices by teacher participants. Prior to the passage of Act 82 in Pennsylvania, the vast majority of educators obtained overall satisfactory ratings, without providing specific information on how to improve. One of the major concerns with any teacher evaluation system is the lack of quality feedback (Mielke & Frontier, 2012). Fostering open feedback between teacher and observer can lead to more effective teaching and provide an opportunity for them to reach their full potential (TNTP, 2012; Wiener & Lundy, 2013). When teachers analytically reflect on their own instructional practices and set improvement goals based on these reflections, teacher motivation and engagement can be enhanced (Mielke & Frontier, 2012). While feedback and self-reflection of instructional practices are integral components of the Formal Observation Model in Pennsylvania's new evaluation system, the presentation of portfolio artifacts before colleagues in the Differentiated Supervision Model can promote "active involvement of participants, encouragement of reflection and self-assessment, and facilitation of collaborative interaction" (Tucker et al., 2003, p. 575).

Although this study provides evidence that teachers' use of a carefully structured portfolio as a reflection tool may result in improved Classroom Instructional Practices, the final Educator Effectiveness ratings of teachers in the Portfolio Mode lacked the discrimination necessary to meet the summative goals for teacher evaluation. The new Pennsylvania teacher evaluation system requires a great investment of time in order for administrators to learn how to fairly, objectively, and reliably evaluate their teachers. If administrators implement the processes mainly to remain compliant, the opportunity to take advantage of the rich discussions about classroom practices embedded in the protocol is lost (Jackson, 2014). Regardless of the method chosen to evaluate teacher effectiveness, unless the process results in the continuous use of best practices of classroom instruction by teachers, improvements in student achievement are unlikely to occur.

A tremendous amount of human and financial resources has been expended to develop teacher evaluation protocols to meet demands for accountability. The potential for these protocols to impact teachers' use of classroom best practices is an important consideration for the educational community as well as for policy-makers. While the national and state focus is on teacher accountability and complex systems to evaluate effective teaching, unless the evaluation process eventually leads to improved teaching practices, improved student learning may not result. As Mielke and Frontier (2012) so eloquently stated: "Only by empowering teachers as the central users of comprehensive teaching frameworks can we ensure that the evaluation system improves teacher effectiveness, rather than merely measuring it" (p. 13).

References

Alvarez, M. E., & Anderson-Ketchmark, C. (2011). Trends and resources: Danielson's framework for teaching. *Children & Schools, 33*(1), 61-63.

Antoniou, P., & Kyriakides, L. (2013). A dynamic integrated approach to teacher professional development: Impact and sustainability of the effects on improving teacher behavior and student outcomes. *Teaching and Teacher Education, 29*(2013), 1-12.

ASCD. (2007). *How to conduct effective high school classroom walk-throughs.* Alexandria, VA: Association for Supervision and Curriculum Development.

Atinello, J. R., Lare, D., & Waters, F. (2006). The value of teacher portfolios for evaluation and professional growth. *NASSP Bulletin, 90*(2), 132-152.

Bambrick-Santoyo, P. (2012). Beyond the scoreboard. *Educational Leadership, 70*(3), 26-30.

Bennett, C. (2012). Will evaluations be fair? *Educational Leadership, 70*(3), 90-91.

Berrill, D. P., & Whalen, C. (2007). "Where are the children?" Personal integrity and reflective teaching portfolios. *Teaching and Teacher Education, 23*(2007), 868-884.

Bill and Melinda Gates Foundation. (2011). *Learning about teaching: Initial findings from the measures of effective teaching project.* Bellevue, WA: Author. Retrieved from www.gatesfoundation.org/college-ready-education/Documents/prelminary-findings-research-paper.pdf

Blake, J., Bachman, J., Frys, M. K., Holbert, P., Ivan, T., & Sellitto, P. (1995). A portfolio-based assessment model for teachers: Encouraging professional growth. *NASSP Bulletin,* October 1995.

Brophy, J. E. (1979). Advances in teacher effectiveness. *Journal of Classroom Interaction, 45*(1), 17-24.

Brophy, J. E., & Evertson, C. (1977). Teacher behaviors and student learning in second and third grades. In G. D. Borich (Ed.), *The appraisal of teaching: Concepts and process* (pp. 79-95). Reading, MA: Addison-Wesley.

Caldas, S. J. (2012). Value-added: The emperor with no clothes. *Educational Leadership, 70*(3), 1-4.

Centra, J. A. (2000). Evaluating the teaching portfolio: A role for colleagues. *New Directions for Teaching and Learning, 83*(Fall 2000), 87-93.

Cooper, H. M., Burger, J. M., & Seymour, G. E. (1979). Classroom context and student ability as influences on teacher perceptions of classroom control. *American Educational Research Journal, 16,* 189-196.

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

Cotton, K. (2003). *Principals and student achievement: What the research says.* Alexandria, VA: Association for Supervision and Curriculum Development.

Council for the Accreditation of Educator Preparation [CAEP]. (2013). *CAEP accreditation standards*. CAEP Commission Recommendations to the CAEP

Board of Directors on Aug., 29, 2013. Retrieved from http://caepnet.org/accreditation/standards/

Danielson, C. (2008). *The handbook for enhancing professional practice.* Alexandria, VA: Association for Supervision and Curriculum Development.

Danielson, C. (2011). *The framework for teaching evaluation instrument*. Princeton, NJ: The Danielson Group. Obtained by request at http://www.danielsongroup.org/downloads.aspx?file=DFfTEI2011.pdf

Danielson, C. (2012). Observing classroom practice. *Educational Leadership 70*(3), 32-37.

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

David, J. L. (2007-2008). What research says about …classroom walk-throughs. *Educational Leadership, 65*(4), 81-82.

Dexter, R. (Spring 2005). Classroom walk-through with reflective feedback: Principals' perceptions of the learning 24/7 classroom walk-through model. *National Forum of Educational Administration and Supervision Journal, 22*(3), 24-39.

DiCarlo**,** M. (2012). How to use value-added measures right. *Educational Leadership, 70*(3), pp. 38-42.

DuFour, R., & Marzano, R. J. (2009). High-leverage strategies for principal leadership. *Educational Leadership, 66*(5), 62-68.

Downey, C. J., Steffy, B. E., English, F. W., Frase, L. E., & Poston, Jr., W. K. (2004). *The three-minute classroom walk-through: Changing school supervisory practice one teacher at a time.* Thousand Oaks, CA: Corwin Press.

Doyle, W. (1977). Paradigms for research on teacher effectiveness. *Review of Research in Education, 5,* 163-198.

Field, A. (2009). *Discovering statistics using SPSS.* Thousand Oaks, CA: Sage Publications.

Fink, E., & Resnick, L. B. (2001). Developing principals as instructional leaders. *Phi Delta Kappan, 82*(8), 598-606.

Fives, H., & Buehl, M. M. (2009). Examining the factor structure of the teachers' sense of efficacy scale. *Journal of Experimental Education, 78*(1), 118-134.

Floden, R. E., & Klinzing, H. G. (1990). What can research on teacher thinking contribute to teacher preparation? A second opinion. *Educational Researcher, 19*(5), 15-20.

Fowler, F. J. (2014). *Survey research methods.* Los Angeles, CA: Sage Publications

Frase, L. E. (2001). *A confirming study of the predictive power of principal classroom visits on efficacy and teacher flow experiences*. Presentation at the Annual meeting of the American Educational Research Association, Seattle, WA.

Frase, L. E., & Hetzel, R. (2002). *School management by wandering around.* Lancaster, PA: Technomic Publishing. (1990 reprinted 2002).

Gabriel, R., & Allington, R. (2012). The MET Project: The wrong 45 million dollar question. *Educational Leadership, 70*(3), 44-49.

Gates, B. (September, 2012). The keys to effective teacher evaluation. *eSchoolNews.* Excerpt from speech given at the Education Commission of the States conference in Atlanta, GA, July 11, 2012. Retrieved from http://www.eschoolnews.com/2012/07/17/bill-gates-the-keys-to-effective-teacher-evaluation/

Gelfer, J. I., Xu, Y., & Perkins, P. G. (2004). Developing portfolios to evaluate teacher performance in early childhood education. *Early Childhood Education Journal 32*(2), 127-132.

Ginsberg, M. B., & Murphy, D. (2002). How walkthroughs open doors. *Educational Leadership, 59*(8), 34-36.

Goe, L. (2013). Can teacher evaluation improve teaching? *Principal Leadership, 13*(7), 24-29.

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

Goldrick, L. (2002). *Improving teacher evaluation to improve teaching quality*. Washington, DC: National Governors Association. Retrieved from http://www.nga.org

Goldwasser, M., & Bach, A. (April, 2007). *What's the difference between an 'A' and an 'A'? High school students' conceptions of good work* [Draft Report]. Paper presented at the annual meeting of American Educational Research Association, Chicago, IL.

Goodwin, B., & Miller, K. (2012). Research says/Use caution with value-added measures. *Educational* Leadership*, 70*(3), 80-81.

Graf, O., & Werlinich, J. (2002). *Observation frustration...Is there another way? The walkthrough observation tool.* Unpublished manuscript, University of Pittsburgh, Pittsburgh, PA: Principals Academy of Western Pennsylvania.

Grant, J. W., & Drafall, L. E. (1991). Teacher effectiveness research: A review and comparison. *Bulletin of the Council for Research in Music Education, 108*, 31-48.

Gray, S. P., & Streshly, W. A. (2008). *From good schools to great schools: What their principals do well.* A Joint Publication by the National Association of Elementary School Principals and Corwin Press, 109-110.

Guskey, T. R. (1987). Context variables that affect measures of teacher efficacy. *The Journal of Educational Research, 81*(1), 41-47.

Haertel, E. (1986). The valid use of student performance measures for teacher evaluation. *Educational Evaluation and Policy Analysis, 8*(1), 45-60.

Hall, P., & Simeral, A. (2008). *Building teachers' capacity for success: A collaborative approach for coaches and school leaders.* Alexandria, VA: Association for Supervision and Curriculum Development.

Hallinger, P., Heck, R. H., & Murphy, J. (2014). Teacher evaluation and school improvement: An analysis of the evidence. *Educational Assessment, Evaluation and Accountability, 26*, 5-28.

Hanushek, E. A., & Rivkin, S. G. (2010). The quality and distribution of teachers under the No Child Left Behind Act. *Journal of Economic Perspectives, 24*(3), 133-150.

Haycock, K. (1998). Good teaching matters: How well-qualified teachers can close the gap. *Thinking K-16, 3*(2). Retrieved from http://www.edtrust.org/sites/edtrust.org/files/publications/files/k16_summer98.pdf

Hazi, H. M., & Rucinski, D. A. (2009). Teacher evaluation as a policy target for improved student learning: A fifty-state review of statute and regulatory action since NCLB. *Education Policy Analysis Archives, 17*(5). Retrieved from http://epaa.asu.edu/epaa/v17n5/

Heneman, H. G., Kimball, S., & Milanowski, A. (2006). The teacher sense of efficacy scale: Validation evidence and behavioral prediction. *Wisconsin Center for Education Research.* Retrieved at http://www.wcer.wisc.edu

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

Holtzapple, E. (2003). Criterion-related validity evidence for a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education, 17*(3), 207-219.

Holzberger, D., Phillipp, A., & Kunter, M. (2013). How teachers' self-efficacy is related to instructional quality: A longitudinal analysis. *Journal of Educational Psychology, 105*(3), 774-786.

Jackson, N. M. (Feb., 2014). The evolving office: As accountability intensifies, the role of secondary principals expands. *District Administration*, 34-38.

Johnson, R. L., McDaniel, II, F., & Willeke, M. J. (2000). Using portfolios in program evaluation: An investigation of interrater reliability. *American Journal of Evaluation, 21*(1), 65-80.

Kachur, D. S., Stout, J. A., & Edwards, C. L. (2010). *Classroom walkthroughs: To improve teaching and learning.* Larchmont, NY: Eye on Education.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Evaluating teacher effectiveness: Can classroom observations identify practices that raise achievement? *Education Next. R*etrieved at www.educationnext.og

Keruskin, T. (2005). *The perceptions of high school principals on student achievement by conducting walkthroughs.* (Unpublished doctoral dissertation). University of Pittsburgh, Pittsburgh, PA.

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

Kimball, S. M., & Milanowski, A. (2009). Examining teacher evaluation validity and leadership decision making within a standards-based evaluation system. *Educational Administration Quarterly, 45*(1), 34-70.

Klassen, R. M., Tze, V. M., Betts, S. M., & Gordon, K. A. (2011). Teacher efficacy research 1998-2009: Signs of progress or unfulfilled promise? *Educational Psychology Review, 23,* 21-43.

Knapper, C., & Wright, W. A. (2001). Using portfolios to document good teaching: Premises, purposes, practices. *New Directions for Teaching and Learning, 88*(Winter 2001), 19-29.

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.

Lane, S., & Horner, C. (2010). *Teacher and principal standards-based evaluation systems.* Document developed under contract #707089 from the Team Pennsylvania Foundation in collaboration with the Pennsylvania Department of Education. Retrieved from http://www.portal.state.pa.us/portal/server.pt/community/educator_effectiveness_project/20903

Lavy, V. (2007). Using performance-based pay to improve the quality of teachers. *The Future of Children, 17* (1), 87-109.

Lawler, P. A. (1991). *The keys to adult learning: Theory and practical strategies.* Philadelphia, PA: Research for Better Schools.

Little, O., Goe, L., & Bell, C. (2009). *A practical guide to evaluating teacher effectiveness*. National Comprehensive Center for Teacher Quality. Retrieved from http://www.gtlcenter.org/sites/default/files/docs/practicalGuide.pdf

Mandell, E. (2006). *Supervisory practices and their effect on teacher's professional growth.* (Unpublished doctoral dissertation). University of Pittsburgh, Pittsburgh, PA.

Marshall, K. (2005). It's time to rethink teacher supervision and evaluation. *Phi Delta Kappan,* 727-735.

Marshall, K. (2012). Fine-tuning teacher evaluation. *Educational Leadership, 70*(3), 50-53.

Martin, P. C. (2011). Selecting one story and hiding others: How AYP chooses the portrayal of a school. *Current Issues in Education, 14*(1). Retrieved from http://cie.asu.edu/

Marzano. R. J. (2012a). Reducing error in teacher observations scores. *Educational Leadership, 70*(3), 82-83.

Marzano, R. J. (2012b). The two purposes of teacher evaluation. *Educational Leadership, 70*(3), 14-19.

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

Marzano, R. J., Waters, T., & McNulty, B. (2005). *School leadership that works: From research to results.* Alexandria, VA: Association for Supervision and Curriculum Development.

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

Measures of Effective Teaching (MET) Project. (2010). *Learning about teaching: Initial findings from the measures of effective teaching project.* Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from www.metproject.org

Measures of Effective Teaching (MET) Project. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from www.metproject.org

Mielke, P., & Frontier, T. (2012. Keeping improvement in mind. *Educational Leadership, 70*(3), 10-13.

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

Moore, Z., & Bond, N. (2002). The use of portfolios for in-service teacher assessment: A case study of foreign language middle-school teachers in Texas. *Foreign Language Annals, 35*(1), 85-92.

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

No Child Left Behind Act of 2001, 20 U.S.C. §6319 (2008).

PACER Research for Action. (2011). Teacher effectiveness: The national picture and Pennsylvania context. *Pennsylvania Clearinghouse for Education Research Issue Brief.* Retrieved from www.researchforaction.org

PA Code. (2013). Chapter 19: Educator Effectiveness Rating Tool. Retrieved from http://www.pacode.com/secure/data/022/chapter19/chap19toc.html

Pallas, A. M. (2012). The fuzzy scarlet letter. *Educational Leadership, 70*(3), 54-57.

Pennsylvania Department of Education (PDE). (2013). *Educator effectiveness system – Differentiated supervision.* Retrieved from http://www.portal.state.pa.us/portal/server.pt/community/educator_effectiveness_project/20903

Peske, H., & Haycock, K. (2006). *Teaching inequality: How poor and minority students are shortchanged on teacher quality.* The Education Trust. Retrieved from http://files.eric.ed.gov/fulltext/ED494820.pdf

Peterson, K. D., Stevens, D., & Mack, C. (2001). Presenting complex teacher evaluation date: Advantages of dossier organization techniques over portfolios. *Journal of Personnel Evaluation in Education, 15*(2), 121-133.

Phillips, V., & Weingarten, R. (2013). Six steps to effective teacher development and evaluation. *eSchool News, 16*(5), 24.

Pickering, D. (2012). Beyond classroom observation. *Educational Leadership, 70*(3). Retrieved from www.ascd.org/publications/educational-leadership/nov12/vol70/num03/Beyond-Classroom-Observations.aspx

Pieczura, M. (2012). Weighing the pros and cons of TAP. *Educational Leadership, 70*(3), 70-71.

Pitler, H., & Goodwin, B. (2009). Classroom walk-throughs: Learning to see the trees and the forest. *The Learning Principal.* National Staff Development Council, 4*(4),* 1, 6-7.

Popham, W. J. (2013). On serving two masters: Formative and summative teacher evaluation. *Principal Leadership, 13*(7), 18-22.

Riggs, I. M., & Sandlin, R. A. (2000). Teaching portfolios for support of teachers' professional growth. *NASSP Bulletin, 84*(618), 22-27.

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

Rothstein, J., & Mathis, W. J. (2013). *Review of two culminating reports from the MET project.* Boulder, CO: National Education Policy Center. Retrieved from http://nepc.colorado.edu/thinktank/review-MET-final-2013

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee value-added assessment system (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*(3), 247-256. Retrieved from www.sas.com/govedu/edu/ed_eval.pdf

Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011). *Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation.* Chicago, IL: Consortium of Chicago School Research at the University of Chicago.

Sawchuk, S. (2013). Combined measures better at gauging teacher effectiveness, study finds. *Education Week, 32*(17), 1-5.

Schachter, R. (2012). Brave new world of teacher evaluation. *District Administration, 48*(10), 43-47.

Scherer, M. (2012). Teachers under the looking glass. *Educational Leadership, 70*(3), 7.

Schmoker, M. (2006). *Results now: How we can achieve unprecedented improvements in teaching and learning.* Alexandria, VA: Association for Supervision and Curriculum Development.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. W. Gagne, & M. Scriven (Eds.), *American Educational Research Association Monograph Series on Curriculum Evaluation: Vol. 1. Perspectives of curriculum evaluation.* Chicago, IL: Rand McNally.

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454-499.

Seo, K. (2012). Lessons from Korea. *Educational Leadership, 70*(3), 75-78.

Simon, M. (2012). A tale of two districts. *Educational Leadership, 70*(3), 58-63.

Stronge, J. H., Richard, H. B., & Catano, N. (2008). *Qualities of effective principals.* Alexandria, VA: Association for Supervision and Curriculum Development.

Tabachnick, B., & Fidell, L. (2013). *Using multivariate statistics*. Boston, MA: Pearson.

The New Teacher Project (TNTP). (2010). *Teacher evaluation 2.0.* Retrieved from The New Teacher Project website at http://tntp.org

The New Teacher Project (TNTP). (2012). *'MET' made simple: Building research-based teacher evaluations.* Retrieved from The New Teacher Project website at http://tntp.org

Tomlinson, C. A. (2012). One to grow on/The evaluation of my dreams. *Educational Leadership, 70*(3), 88-89.

Trochim, W. M., & Donnelly, J. P. (2008). *The research methods knowledge base.* Mason, OH: Cengage Learning.

Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education, 17*, 783-805.

Tucker, P. D., Stronge, J. H., Gareis, C. R., & Beers, C. S. (2003). The efficacy of portfolios for teacher evaluation and professional development: Do they make a difference? *Educational Administration Quarterly, 39*(5), 572-602.

United States Department of Education (US DOE). (2011). *Race to the top fund: States' applications for phase 3.* Retrieved from http://www2.ed.gov/programs/racetothetop/phase3-applications/index.html

Valli, L., & Buese, D. (2007). The changing roles of teachers in an era of high-stakes accountability. *American Educational Research Journal, 44*(3), 519-558.

Van der Schaaf, M. F., Stokking, K. M., & Verloop, N. (2005). Cognitive representations in raters' assessment of teacher portfolios. *Studies in Educational Evaluation 31*, 27-55.

Veir, C., & Dagley, D. (2002). Legal issues in teacher evaluation legislation: A study of state statutory provisions. *B.Y.U. Education and Law Journal, 2002*(1) [online]. Retrieved from http://www.law2.byu.edu/jel/v2002_1/Veir1.htm

Wade, R. C., & Yarbrough, D. B. (1996). Portfolios: A tool for reflective thinking in teacher education? *Teaching & Teacher Education, 12*(1), 63-79.

Weber, C. (2012). The balancing acts of teacher evaluation. *Educational Leadership, 70*(3). Retrieved from http://www.ascd.org/publications/educational-leadership/nov12/vol70/num03/The-Balancing-Acts-of-Teacher-Evaluation.aspx

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

Westbury, M. (1988). The science and the art of teacher effectiveness: An examination of two research traditions. *Canadian Journal of Education, 13*(1), 138-161.

Westerberg, T. R. (2006). Feedback for teachers: Focused, specific, and constructive. *Principal Leadership, 13*(7), 30-33.

Wiener, R., & Lundy, K. (2013). Evaluating evaluations: Using teacher surveys to strengthen implementation. *The Aspen Institute Education and Society Program.*

Wilkerson, D. J., Manatt, R. P., Rogers, M. A., & Maughan, R. (2000). Validation of student, principal, and self-ratings in 360^0 Feedback® for teacher evaluation. *Journal of Personnel Evaluation in Education, 14*(2), 179-192.

Whitaker, T., & Zoul, J. (2008). *The 4 core factors for school success.* Larchmont, NY: Eye on Education.

Whitney, J., Leonard, M., Leonard, W., Camelio, M., & Camelio, V. (2005). *Seek balance, connect with others, and reach all students: High school students describe a moral imperative for teachers*. University of North Carolina Press: The High School Journal.

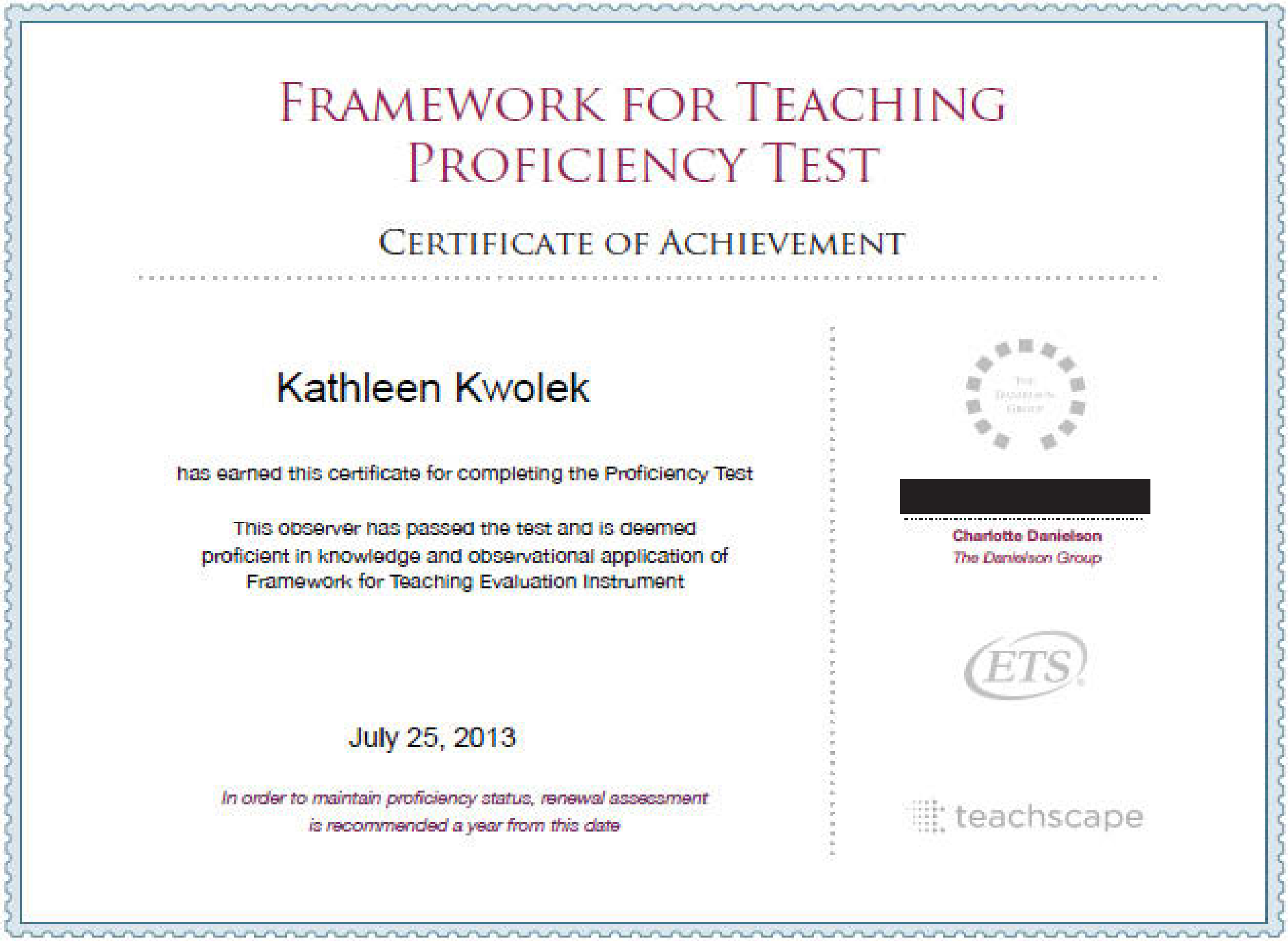
Wolf, K., Lichtenstein, G., Bartlett, E., & Hartman, D. (1996). Professional development and teaching portfolios: The Douglas County outstanding teacher program. *Journal of Personnel Evaluation in Education, 10*, 279-286.

Woolfolk, A. E., Rosoff, B., & Hoy, W. K. (1990). Teachers' sense of efficacy and their beliefs about managing students. *Teaching & Teacher Education, 6*(2), 137-148.

Xu, J. (2003). Promoting school-centered professional development through teaching portfolios. *Journal of Teacher Education, 54*(4), 347-361.

Youngs, P. (2013). *Using teacher evaluation reform and professional development to support common core assessments.* Center for American Progress, www.americanprogress.org

Appendix C



FRAMEWORK FOR TEACHING PROFICIENCY TEST

CERTIFICATE OF ACHIEVEMENT

Kathleen Kwolek

has earned this certificate for completing the Proficiency Test

This observer has passed the test and is deemed proficient in knowledge and observational application of Framework for Teaching Evaluation Instrument

July 25, 2013

In order to maintain proficiency status, renewal assessment is recommended a year from this date

[redacted]

Charlotte Danielson
The Danielson Group

ETS

teachscape

Appendix E

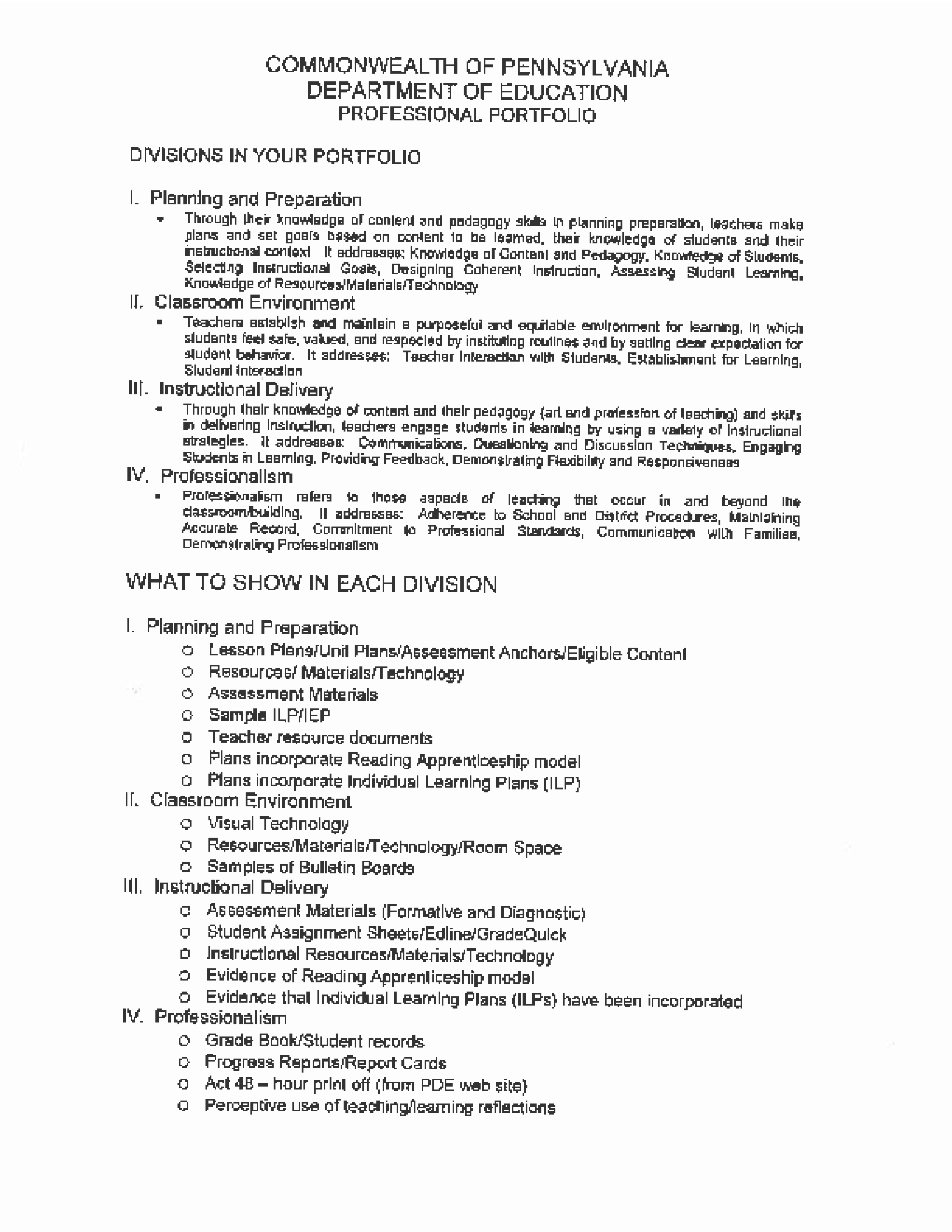
Focus Area	Failing (0)	Needs Improvement (1)	Proficient (2)	Distinguished (3)
What are the Artifacts? Selection of Artifacts, Knowledge of Teacher Effectiveness Domains and Components, Curriculum – Standards Aligned	No artifacts were provided. Teacher displays no understanding of the concepts contained in the Domains. Teacher does not identify standards and/or how Artifact meets the standards.	Artifacts do not clearly reflect the Domains. Teacher displays a minimal understanding of the concepts contained in the Domains. Teacher identifies standards, but has a weak ability to explain relationship to standards.	The artifacts are related to the concepts contained in the Domains. Teacher displays a competent understanding of the concept contained in the Domains. Teacher identifies standards and explains how the artifacts meet the standards.	Multiple artifacts are directly related to the Domains. Teacher displays an extensive understanding of the concepts of the Domains. Teacher demonstrates a thorough understanding of purpose of artifacts in relation to course and standards.
Why did you choose the Artifacts? Reflections, Goals, Feedback	Teacher displays no evidence of reflection specific to the Domains. No goals for professional improvement. Teacher does not seek feedback for professional improvement.	Reflections are a summary of the activity or artifacts. Goals for professional improvement are not clearly articulated. Teacher seeks out feedback for professional improvement, but does not implement recommended strategies.	Reflections are clear and directly related to the Domains. Goals for professional improvement are articulated and teacher seeks out feedback from administration and colleagues and attempts to implements suggested strategies.	Reflections show evidence of thoughtful study related to the Domains, citing specific examples. There is a plan of action for professional improvement, including evidence of seeking out feedback from administration and colleagues and implementation of recommended strategies.
How do the Artifacts impact teaching and learning?	Teacher does not relate the artifacts to professional learning or student performance. Relation for teaching and learning is not present.	Teacher attempts to relate artifacts to professional learning or student performance, but the connection is not clear. Relation for teaching and learning is minimal.	Teacher clearly relates artifacts to professional learning and student performance. Relation for teaching and learning is accurate.	Teacher successfully relates artifacts to professional learning and uses evidence to show correlation to student performance. Relation for teaching and learning is fully accurate.

Portfolio Evaluation June, 2014

Evaluator: __________________ Teacher Name: ___________________

0 = Failing; 1 = Needs Improvement, 2 = Proficient; 3 = Distinguished

	Domain 1		Domain 2		Domain 3		Domain 4	
What are the artifacts? Component	–	–	–	–	–	–	–	–
Related to the concepts in the Domain	–	–	–	–	–	–	–	–
T displays a competent understanding of concept	–	–	–	–	–	–	–	–
T explains how artifacts meets the standards	–	–	–	–	–	–	–	–
Why did you choose the artifact?								
Reflections are clear and related to Domain	–	–	–	–	–	–	–	–
Goals for prof development are provided	–	–	–	–	–	–	–	–
T seeks out feedback and attempts to implement	–	–	–	–	–	–	–	–
How do the artifacts impact teaching and learning?								
T clearly relates artifacts to prof learning	–	–	–	–	–	–	–	–
T relates artifacts to student performance	–	–	–	–	–	–	–	–
Relation for teaching and learning is accurate	–	–	–	–	–	–	–	–



COMMONWEALTH OF PENNSYLVANIA
DEPARTMENT OF EDUCATION
PROFESSIONAL PORTFOLIO

DIVISIONS IN YOUR PORTFOLIO

I. Planning and Preparation

- Through their knowledge of content and pedagogy skills in planning preparation, teachers make plans and set goals based on content to be learned, their knowledge of students and their instructional context It addresses: Knowledge of Content and Pedagogy, Knowledge of Students, Selecting Instructional Goals, Designing Coherent Instruction, Assessing Student Learning, Knowledge of Resources/Materials/Technology

II. Classroom Environment

- Teachers establish and maintain a purposeful and equitable environment for learning, in which students feel safe, valued, and respected by instituting routines and by setting clear expectation for student behavior. It addresses: Teacher Interaction with Students, Establishment for Learning, Student interaction

III. Instructional Delivery

- Through their knowledge of content and their pedagogy (art and profession of teaching) and skills in delivering instruction, teachers engage students in learning by using a variety of instructional strategies. It addresses: Communications, Questioning and Discussion Techniques, Engaging Students in Learning, Providing Feedback, Demonstrating Flexibility and Responsiveness

IV. Professionalism

- Professionalism refers to those aspects of teaching that occur in and beyond the classroom/building. It addresses: Adherence to School and District Procedures, Maintaining Accurate Record, Commitment to Professional Standards, Communication with Families, Demonstrating Professionalism

WHAT TO SHOW IN EACH DIVISION

I. Planning and Preparation

- ○ Lesson Plans/Unit Plans/Assessment Anchors/Eligible Content
- ○ Resources/ Materials/Technology
- ○ Assessment Materials
- ○ Sample ILP/IEP
- ○ Teacher resource documents
- ○ Plans incorporate Reading Apprenticeship model
- ○ Plans incorporate Individual Learning Plans (ILP)

II. Classroom Environment

- ○ Visual Technology
- ○ Resources/Materials/Technology/Room Space
- ○ Samples of Bulletin Boards

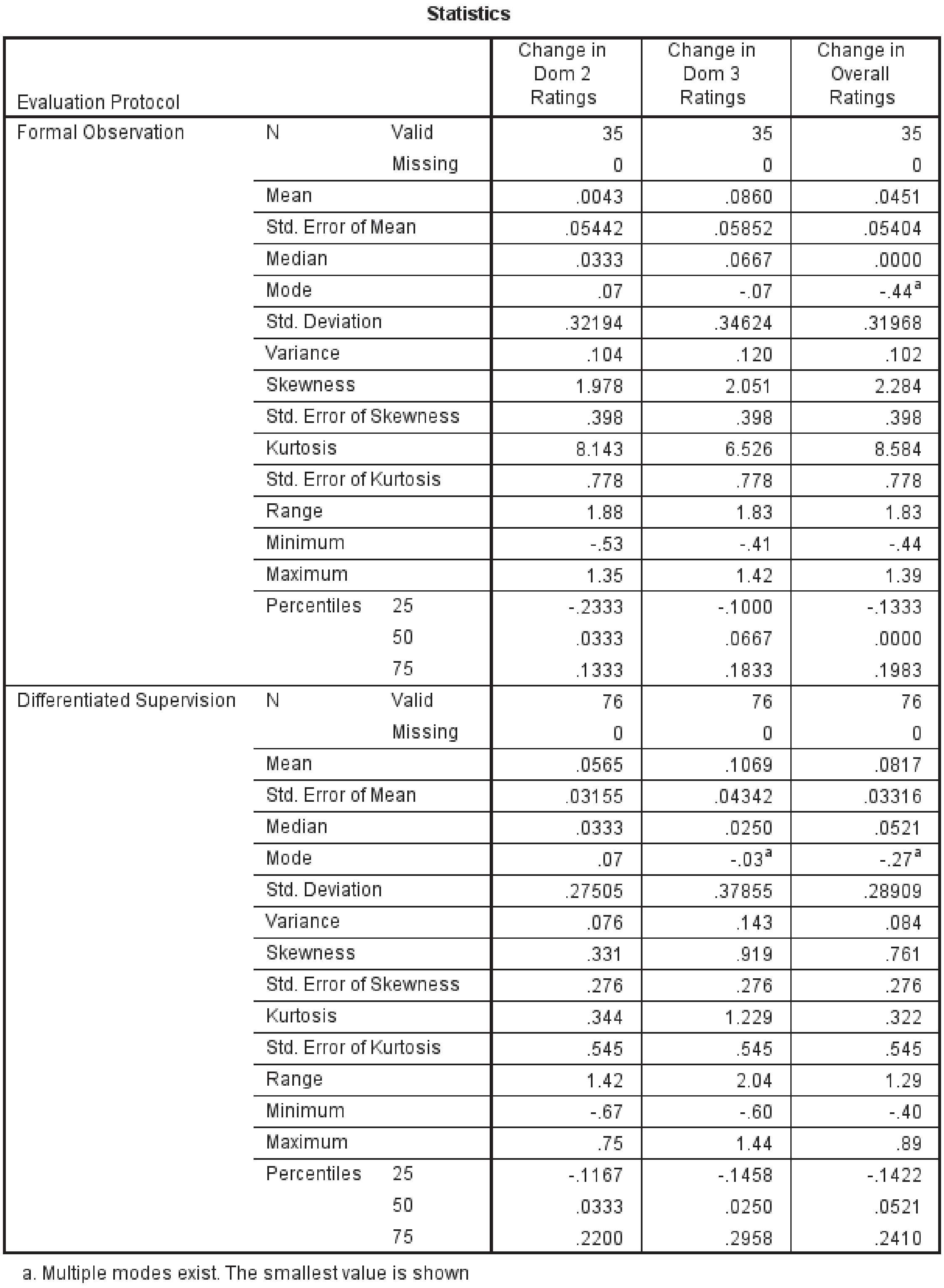
III. Instructional Delivery

- ○ Assessment Materials (Formative and Diagnostic)
- ○ Student Assignment Sheets/Edline/GradeQuick
- ○ Instructional Resources/Materials/Technology
- ○ Evidence of Reading Apprenticeship model
- ○ Evidence that Individual Learning Plans (ILPs) have been incorporated

IV. Professionalism

- ○ Grade Book/Student records
- ○ Progress Reports/Report Cards
- ○ Act 48 – hour print off (from PDE web site)
- ○ Perceptive use of teaching/learning reflections

Appendix F



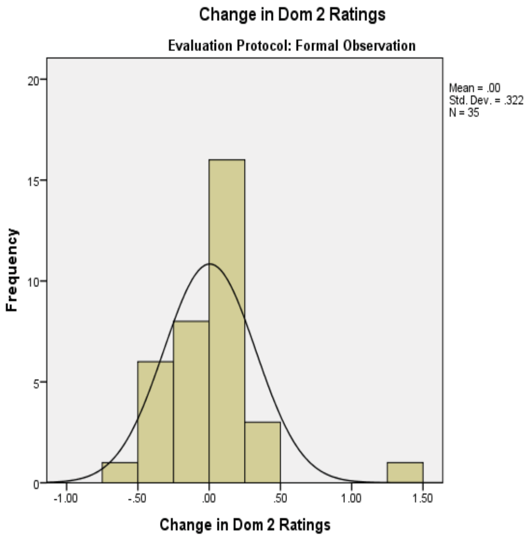
Statistics

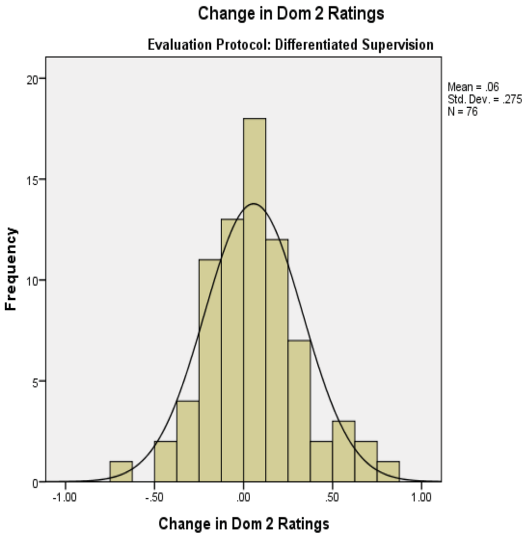
Evaluation Protocol			Change in Dom 2 Ratings	Change in Dom 3 Ratings	Change in Overall Ratings
Formal Observation	N	Valid	35	35	35
		Missing	0	0	0
	Mean		.0043	.0860	.0451
	Std. Error of Mean		.05442	.05852	.05404
	Median		.0333	.0667	.0000
	Mode		.07	-.07	-.44[a]
	Std. Deviation		.32194	.34624	.31968
	Variance		.104	.120	.102
	Skewness		1.978	2.051	2.284
	Std. Error of Skewness		.398	.398	.398
	Kurtosis		8.143	6.526	8.584
	Std. Error of Kurtosis		.778	.778	.778
	Range		1.88	1.83	1.83
	Minimum		-.53	-.41	-.44
	Maximum		1.35	1.42	1.39
	Percentiles	25	-.2333	-.1000	-.1333
		50	.0333	.0667	.0000
		75	.1333	.1833	.1983
Differentiated Supervision	N	Valid	76	76	76
		Missing	0	0	0
	Mean		.0565	.1069	.0817
	Std. Error of Mean		.03155	.04342	.03316
	Median		.0333	.0250	.0521
	Mode		.07	-.03[a]	-.27[a]
	Std. Deviation		.27505	.37855	.28909
	Variance		.076	.143	.084
	Skewness		.331	.919	.761
	Std. Error of Skewness		.276	.276	.276
	Kurtosis		.344	1.229	.322
	Std. Error of Kurtosis		.545	.545	.545
	Range		1.42	2.04	1.29
	Minimum		-.67	-.60	-.40
	Maximum		.75	1.44	.89
	Percentiles	25	-.1167	-.1458	-.1422
		50	.0333	.0250	.0521
		75	.2200	.2958	.2410

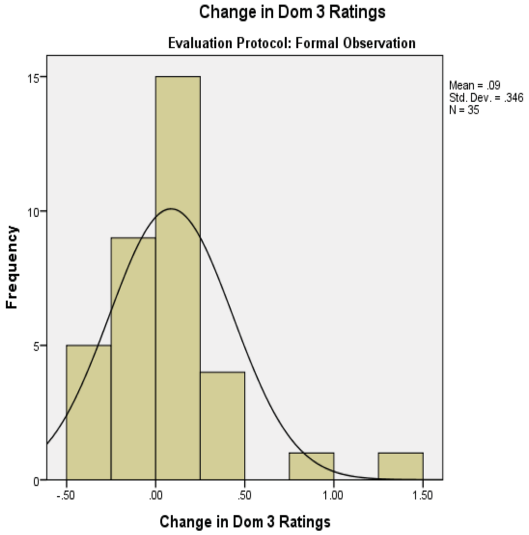
a. Multiple modes exist. The smallest value is shown

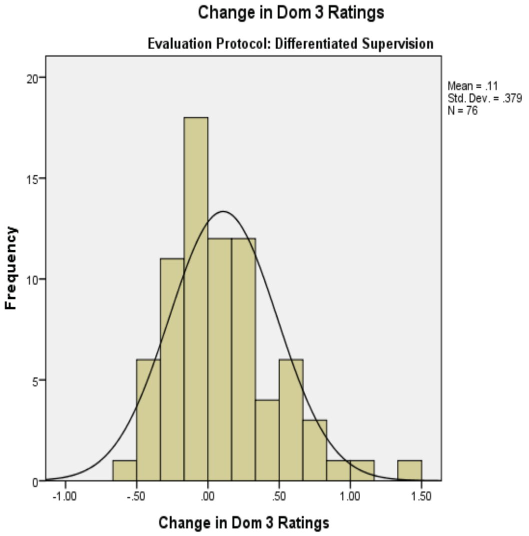
Changes in Participants' Observation Ratings

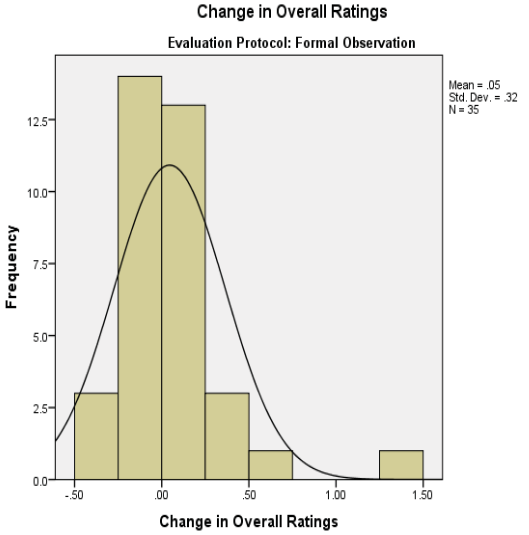
	N	M	SD	Skewness	Kurtosis
Changes in Domain 2 Ratings					
Formal Observation	35	.004	.32	1.98	8.14
Differentiated Supervision	76	.06	.28	.33	.34
Changes in Domain 3 Ratings					
Formal Observation	35	.09	.35	2.05	6.53
Differentiated Supervision	76	.11	.38	.92	1.23
Changes in Overall Ratings					
Formal Observation	35	.045	.32	2.28	8.58
Differentiated Supervision	76	.08	.29	.76	.32

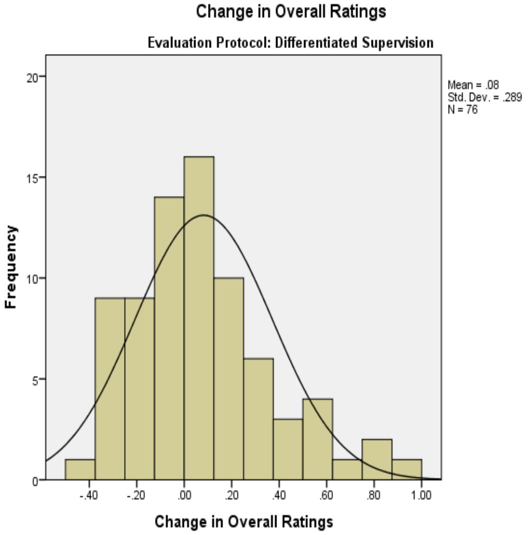

Change in Dom 2 Ratings
Evaluation Protocol: Formal Observation
Mean = .00
Std. Dev. = .322
N = 35
Frequency
20
15
10
5
0
-1.00
-.50
.00
.50
1.00
1.50
Change in Dom 2 Ratings

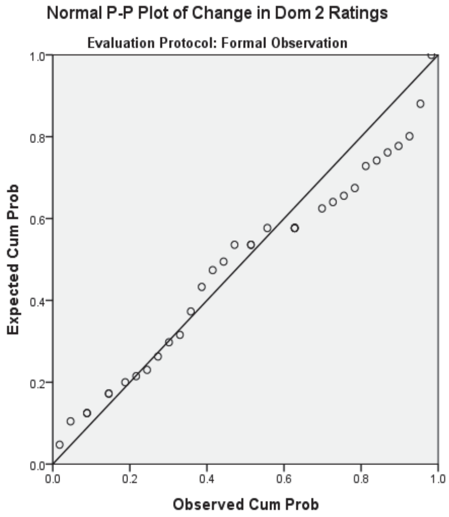

Change in Dom 2 Ratings
Evaluation Protocol: Differentiated Supervision
Mean = .06
Std. Dev. = .275
N = 76
Frequency
20
15
10
5
0
-1.00
-.50
.00
.50
1.00
Change in Dom 2 Ratings

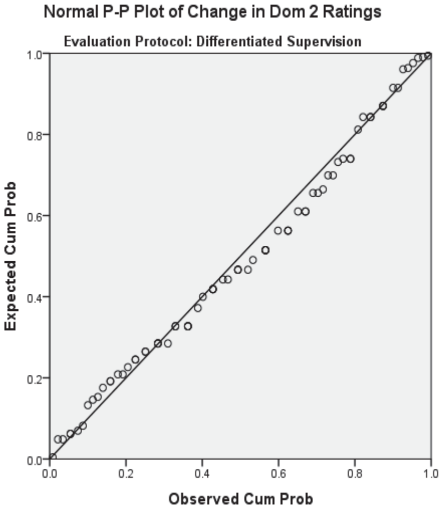

Change in Dom 3 Ratings
Evaluation Protocol: Formal Observation
Mean = .09
Std. Dev. = .346
N = 35
Frequency
15
10
5
0
-.50
.00
.50
1.00
1.50
Change in Dom 3 Ratings

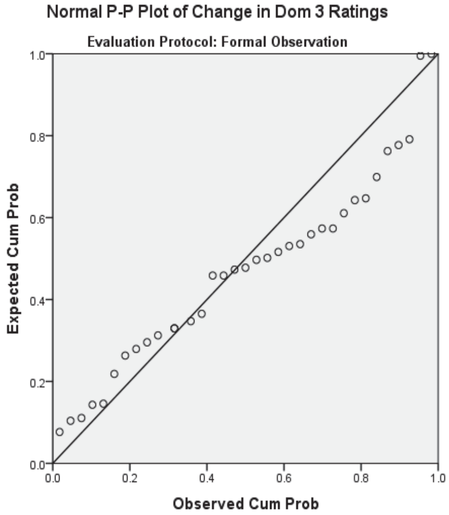

Change in Dom 3 Ratings
Evaluation Protocol: Differentiated Supervision
Mean = .11
Std. Dev. = .379
N = 76
Frequency
20
15
10
5
0
-1.00
-.50
.00
.50
1.00
1.50
Change in Dom 3 Ratings

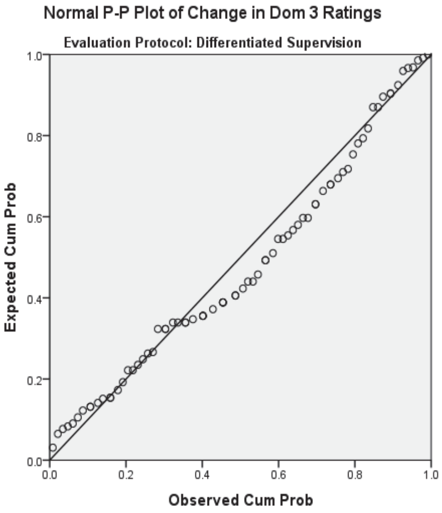

Change in Overall Ratings
Evaluation Protocol: Formal Observation
Mean = .05
Std. Dev. = .32
N = 35
Frequency
12.5
10.0
7.5
5.0
2.5
0.0
-.50
.00
.50
1.00
1.50
Change in Overall Ratings

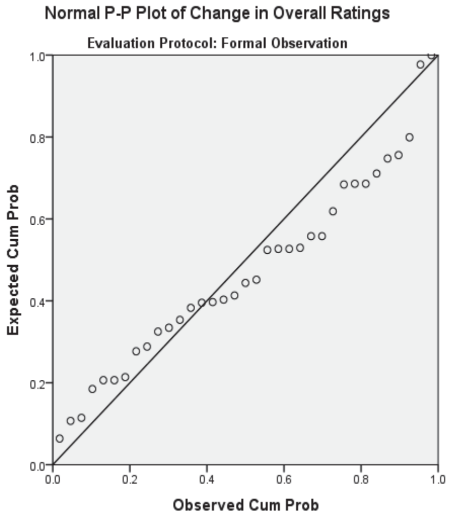

Change in Overall Ratings
Evaluation Protocol: Differentiated Supervision
Mean = .08
Std. Dev. = .289
N = 76
Frequency
20
15
10
5
0
-.40
-.20
.00
.20
.40
.60
.80
1.00
Change in Overall Ratings

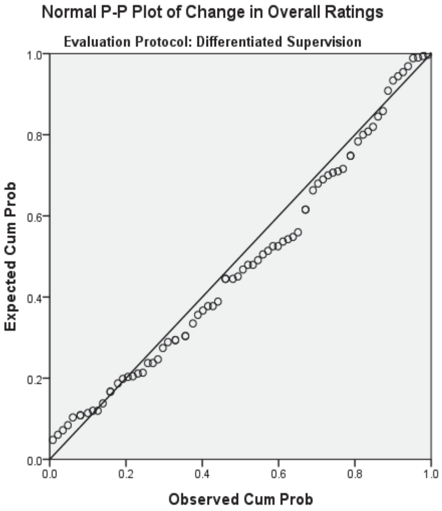

Normal P-P Plot of Change in Dom 2 Ratings
Evaluation Protocol: Formal Observation
Expected Cum Prob
Observed Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0


Normal P-P Plot of Change in Dom 2 Ratings
Evaluation Protocol: Differentiated Supervision
Expected Cum Prob
Observed Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0

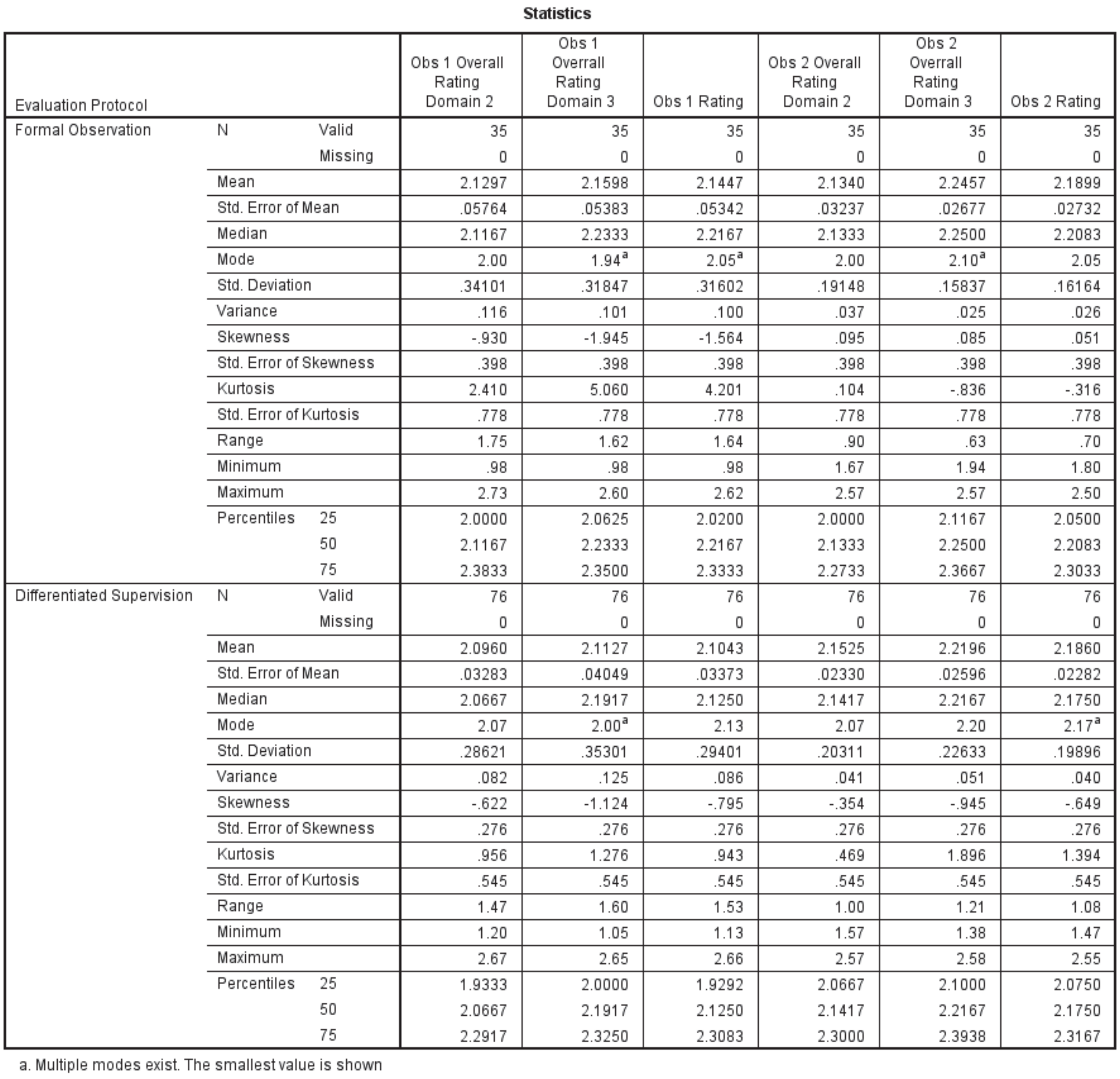

Normal P-P Plot of Change in Dom 3 Ratings
Evaluation Protocol: Formal Observation
Expected Cum Prob
Observed Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0


Normal P-P Plot of Change in Dom 3 Ratings
Evaluation Protocol: Differentiated Supervision
Expected Cum Prob
Observed Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0


Normal P-P Plot of Change in Overall Ratings
Evaluation Protocol: Formal Observation
Expected Cum Prob
Observed Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0


Normal P-P Plot of Change in Overall Ratings
Evaluation Protocol: Differentiated Supervision
Expected Cum Prob
Observed Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0

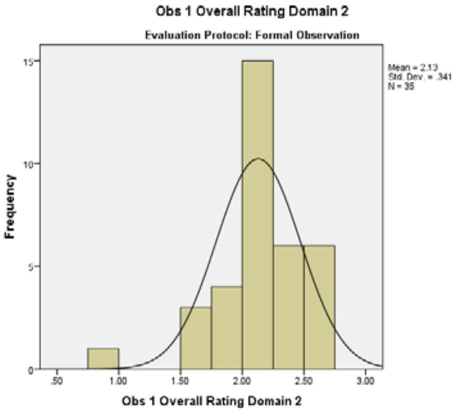
Appendix G

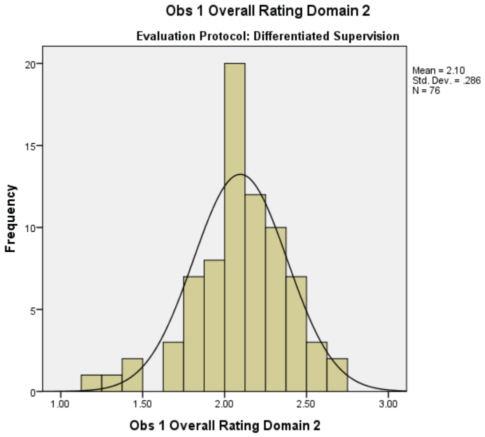


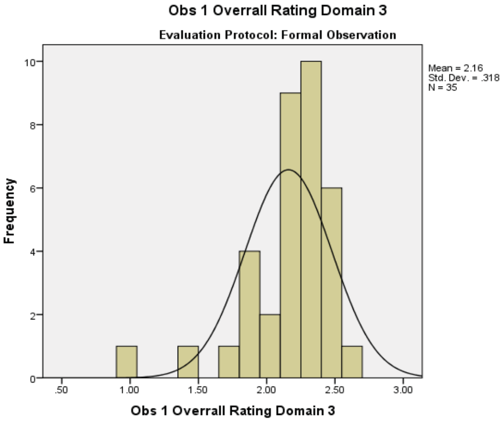
Statistics

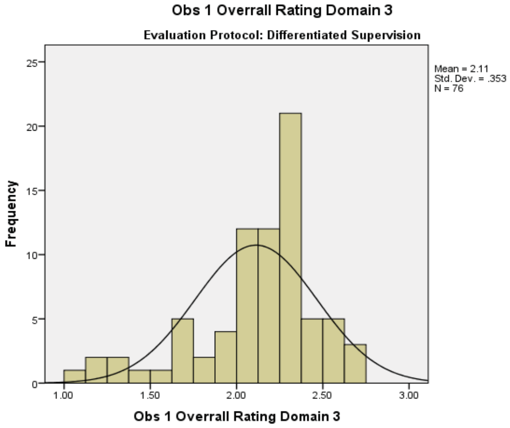
Evaluation Protocol			Obs 1 Overall Rating Domain 2	Obs 1 Overall Rating Domain 3	Obs 1 Rating	Obs 2 Overall Rating Domain 2	Obs 2 Overall Rating Domain 3	Obs 2 Rating
Formal Observation	N	Valid	35	35	35	35	35	35
		Missing	0	0	0	0	0	0
	Mean		2.1297	2.1598	2.1447	2.1340	2.2457	2.1899
	Std. Error of Mean		.05764	.05383	.05342	.03237	.02677	.02732
	Median		2.1167	2.2333	2.2167	2.1333	2.2500	2.2083
	Mode		2.00	1.94[a]	2.05[a]	2.00	2.10[a]	2.05
	Std. Deviation		.34101	.31847	.31602	.19148	.15837	.16164
	Variance		.116	.101	.100	.037	.025	.026
	Skewness		-.930	-1.945	-1.564	.095	.085	.051
	Std. Error of Skewness		.398	.398	.398	.398	.398	.398
	Kurtosis		2.410	5.060	4.201	.104	-.836	-.316
	Std. Error of Kurtosis		.778	.778	.778	.778	.778	.778
	Range		1.75	1.62	1.64	.90	.63	.70
	Minimum		.98	.98	.98	1.67	1.94	1.80
	Maximum		2.73	2.60	2.62	2.57	2.57	2.50
	Percentiles	25	2.0000	2.0625	2.0200	2.0000	2.1167	2.0500
		50	2.1167	2.2333	2.2167	2.1333	2.2500	2.2083
		75	2.3833	2.3500	2.3333	2.2733	2.3667	2.3033
Differentiated Supervision	N	Valid	76	76	76	76	76	76
		Missing	0	0	0	0	0	0
	Mean		2.0960	2.1127	2.1043	2.1525	2.2196	2.1860
	Std. Error of Mean		.03283	.04049	.03373	.02330	.02596	.02282
	Median		2.0667	2.1917	2.1250	2.1417	2.2167	2.1750
	Mode		2.07	2.00[a]	2.13	2.07	2.20	2.17[a]
	Std. Deviation		.28621	.35301	.29401	.20311	.22633	.19896
	Variance		.082	.125	.086	.041	.051	.040
	Skewness		-.622	-1.124	-.795	-.354	-.945	-.649
	Std. Error of Skewness		.276	.276	.276	.276	.276	.276
	Kurtosis		.956	1.276	.943	.469	1.896	1.394
	Std. Error of Kurtosis		.545	.545	.545	.545	.545	.545
	Range		1.47	1.60	1.53	1.00	1.21	1.08
	Minimum		1.20	1.05	1.13	1.57	1.38	1.47
	Maximum		2.67	2.65	2.66	2.57	2.58	2.55
	Percentiles	25	1.9333	2.0000	1.9292	2.0667	2.1000	2.0750
		50	2.0667	2.1917	2.1250	2.1417	2.2167	2.1750
		75	2.2917	2.3250	2.3083	2.3000	2.3938	2.3167

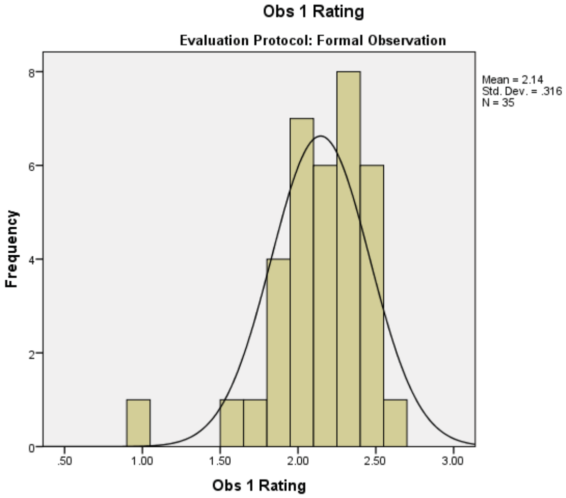
a. Multiple modes exist. The smallest value is shown

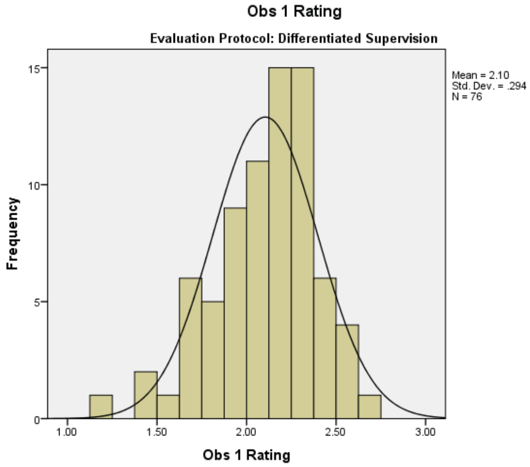

Obs 1 Overall Rating Domain 2
Evaluation Protocol: Formal Observation
Mean = 2.13
Std. Dev. = .341
N = 35
Frequency
15
10
5
0
.50
1.00
1.50
2.00
2.50
3.00
Obs 1 Overall Rating Domain 2

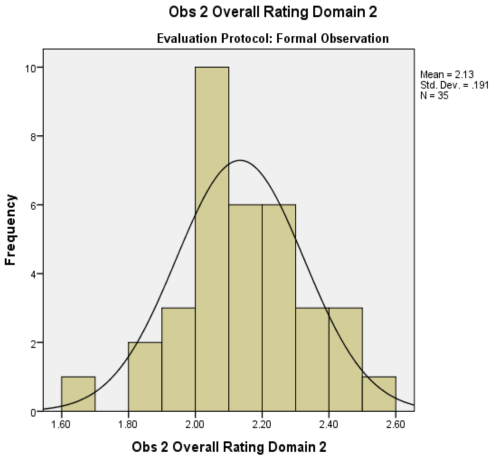

Obs 1 Overall Rating Domain 2
Evaluation Protocol: Differentiated Supervision
Mean = 2.10
Std. Dev. = .286
N = 76
Frequency
20
15
10
5
0
1.00
1.50
2.00
2.50
3.00
Obs 1 Overall Rating Domain 2

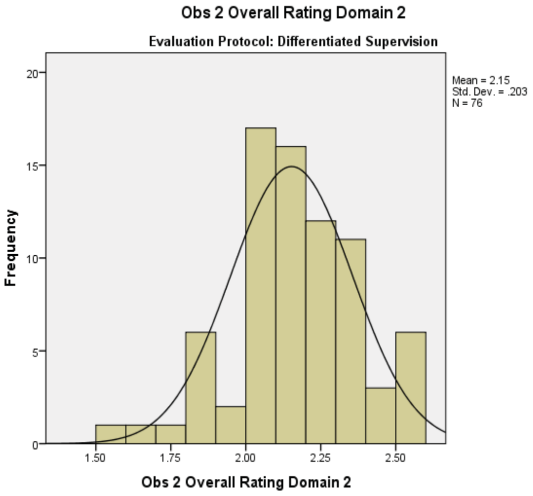

Obs 1 Overall Rating Domain 3
Evaluation Protocol: Formal Observation
Mean = 2.16
Std. Dev. = .318
N = 35
Frequency
10
8
6
4
2
0
.50
1.00
1.50
2.00
2.50
3.00
Obs 1 Overall Rating Domain 3

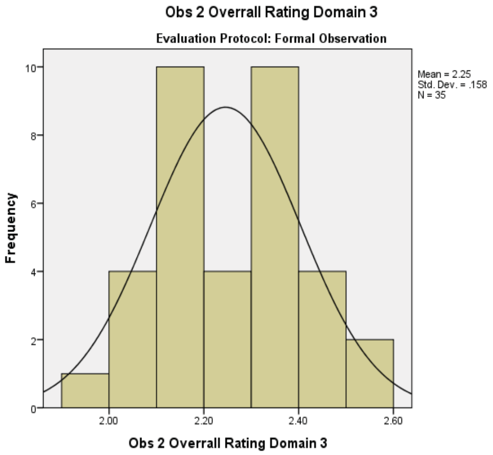

Obs 1 Overall Rating Domain 3
Evaluation Protocol: Differentiated Supervision
Mean = 2.11
Std. Dev. = .353
N = 76
Frequency
25
20
15
10
5
0
1.00
1.50
2.00
2.50
3.00
Obs 1 Overall Rating Domain 3

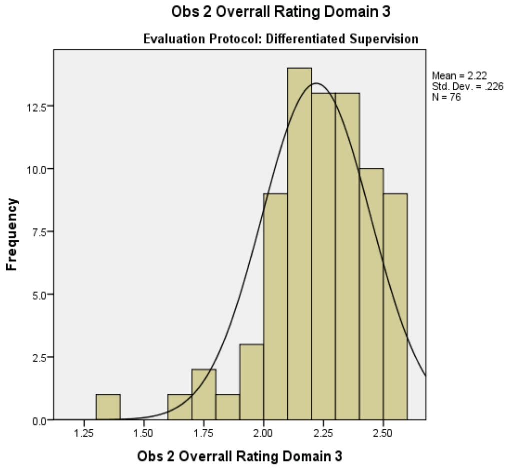

Obs 1 Rating
Evaluation Protocol: Formal Observation
Mean = 2.14
Std. Dev. = .316
N = 35
Frequency
8
6
4
2
0
.50
1.00
1.50
2.00
2.50
3.00
Obs 1 Rating

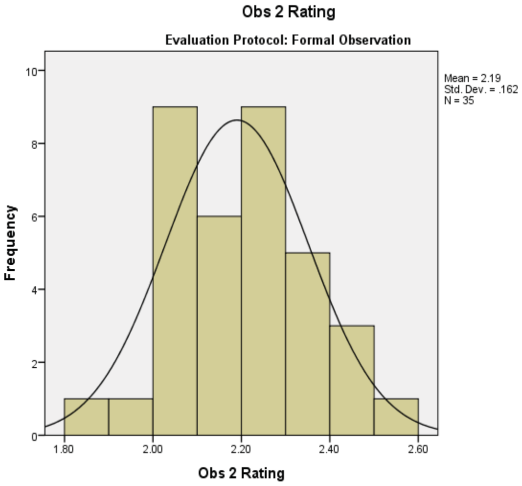

Obs 1 Rating
Evaluation Protocol: Differentiated Supervision
Mean = 2.10
Std. Dev. = .294
N = 76
Frequency
15
10
5
0
1.00
1.50
2.00
2.50
3.00
Obs 1 Rating

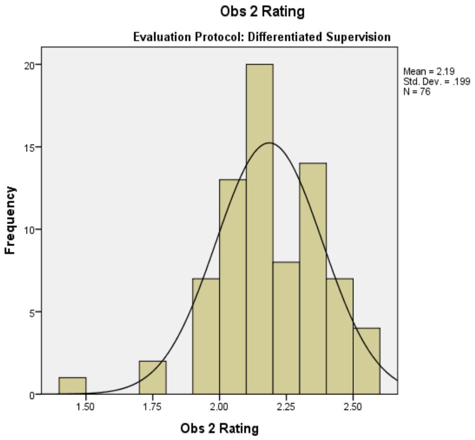

Obs 2 Overall Rating Domain 2
Evaluation Protocol: Formal Observation
Frequency
0
2
4
6
8
10
1.60
1.80
2.00
2.20
2.40
2.60
Obs 2 Overall Rating Domain 2
Mean = 2.13
Std. Dev. = .191
N = 35

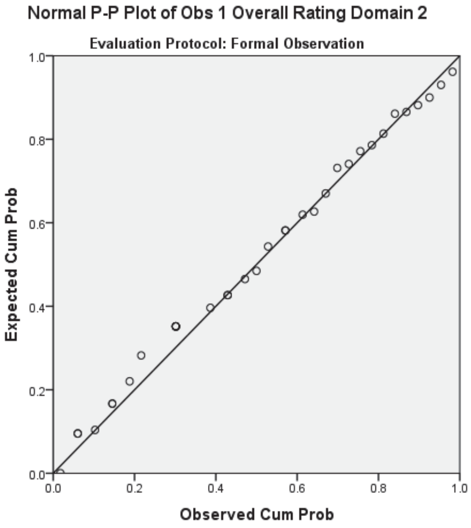

Obs 2 Overall Rating Domain 2
Evaluation Protocol: Differentiated Supervision
Frequency
0
5
10
15
20
1.50
1.75
2.00
2.25
2.50
Obs 2 Overall Rating Domain 2
Mean = 2.15
Std. Dev. = .203
N = 76

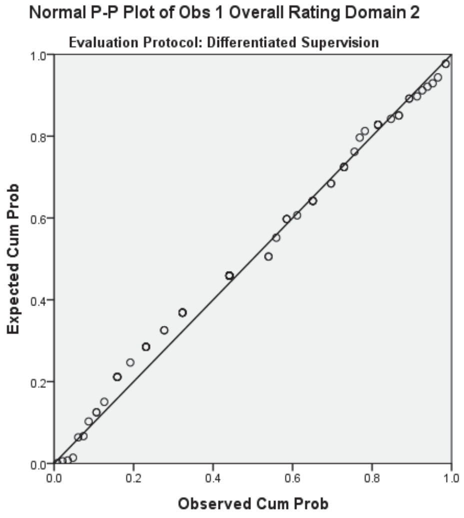

Obs 2 Overrall Rating Domain 3
Evaluation Protocol: Formal Observation
Frequency
0
2
4
6
8
10
2.00
2.20
2.40
2.60
Obs 2 Overrall Rating Domain 3
Mean = 2.25
Std. Dev. = .158
N = 35

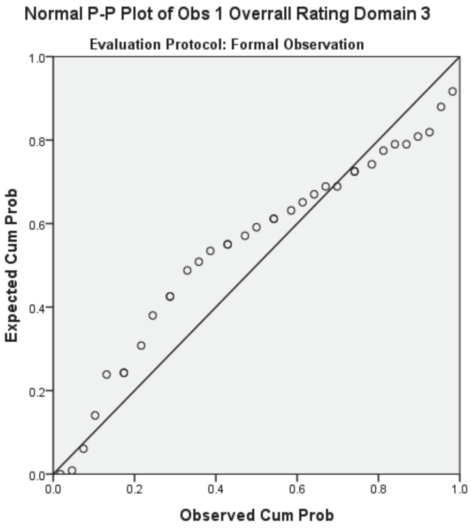

Obs 2 Overrall Rating Domain 3
Evaluation Protocol: Differentiated Supervision
Frequency
0.0
2.5
5.0
7.5
10.0
12.5
1.25
1.50
1.75
2.00
2.25
2.50
Obs 2 Overrall Rating Domain 3
Mean = 2.22
Std. Dev. = .226
N = 76

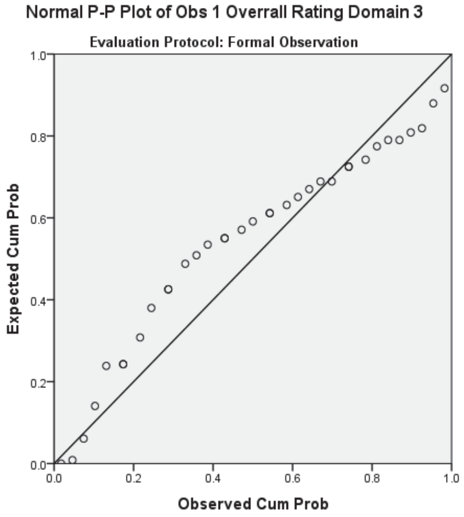

Obs 2 Rating
Evaluation Protocol: Formal Observation
Frequency
0
2
4
6
8
10
1.80
2.00
2.20
2.40
2.60
Obs 2 Rating
Mean = 2.19
Std. Dev. = .162
N = 35

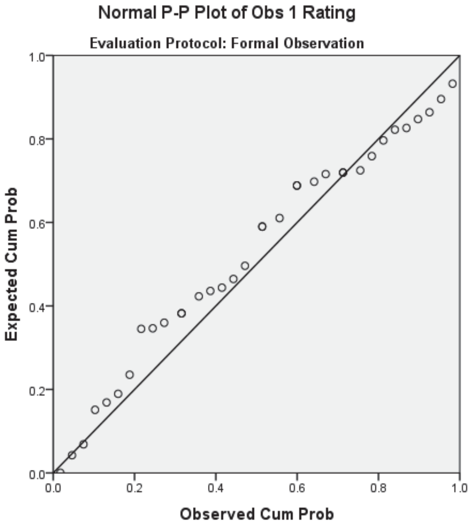

Obs 2 Rating
Evaluation Protocol: Differentiated Supervision
Frequency
0
5
10
15
20
1.50
1.75
2.00
2.25
2.50
Obs 2 Rating
Mean = 2.19
Std. Dev. = .199
N = 76

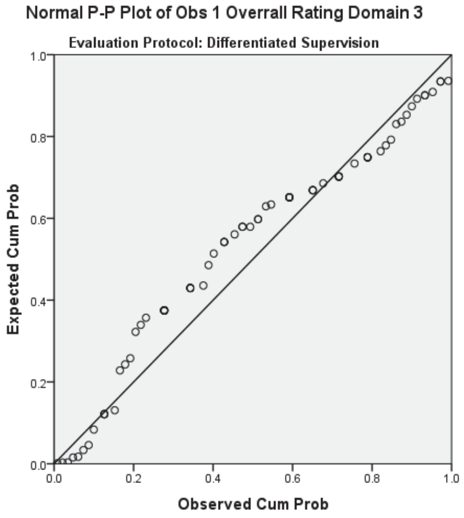

Normal P-P Plot of Obs 1 Overall Rating Domain 2
Evaluation Protocol: Formal Observation
Expected Cum Prob
Observed Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0

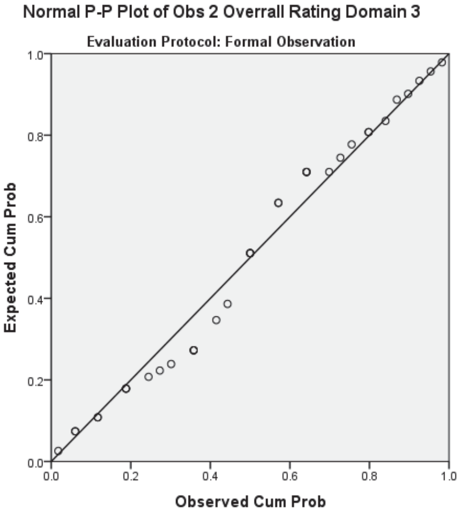

Normal P-P Plot of Obs 1 Overall Rating Domain 2
Evaluation Protocol: Differentiated Supervision
Expected Cum Prob
Observed Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0

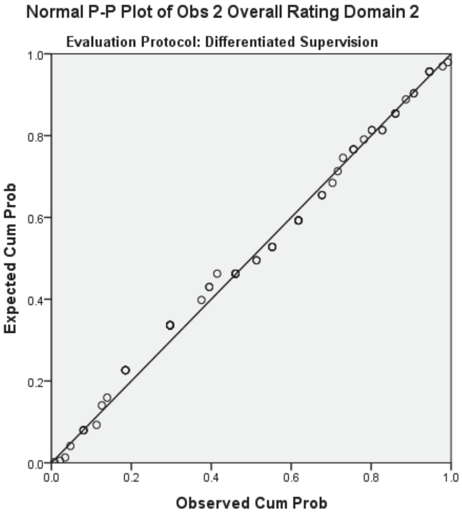

Normal P-P Plot of Obs 1 Overrall Rating Domain 3
Evaluation Protocol: Formal Observation
Expected Cum Prob
Observed Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0

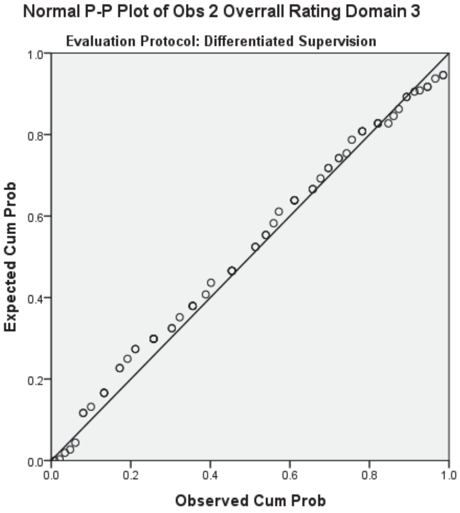

Normal P-P Plot of Obs 1 Overrall Rating Domain 3
Evaluation Protocol: Formal Observation
Expected Cum Prob
Observed Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0

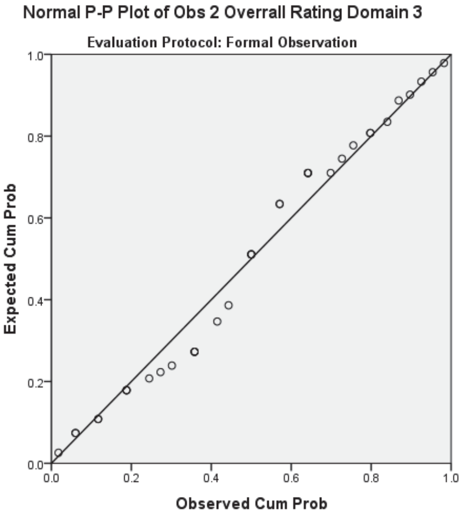

Normal P-P Plot of Obs 1 Rating
Evaluation Protocol: Formal Observation
Expected Cum Prob
Observed Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0

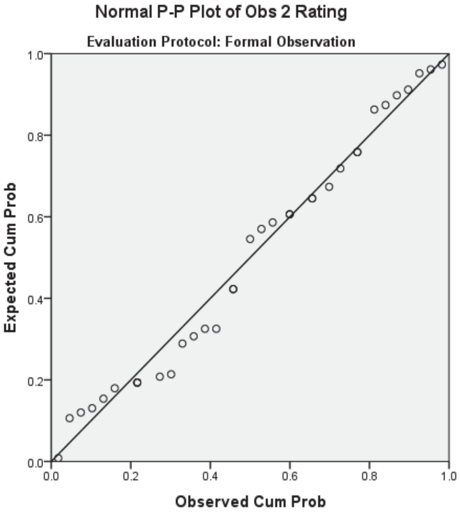

Normal P-P Plot of Obs 1 Overrall Rating Domain 3
Evaluation Protocol: Differentiated Supervision
Expected Cum Prob
Observed Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0

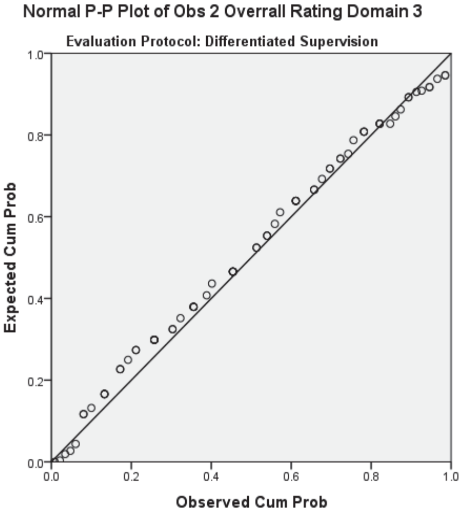

Normal P-P Plot of Obs 2 Overall Rating Domain 3
Evaluation Protocol: Formal Observation
Expected Cum Prob
Observed Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0


Normal P-P Plot of Obs 2 Overall Rating Domain 2
Evaluation Protocol: Differentiated Supervision
Expected Cum Prob
Observed Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0

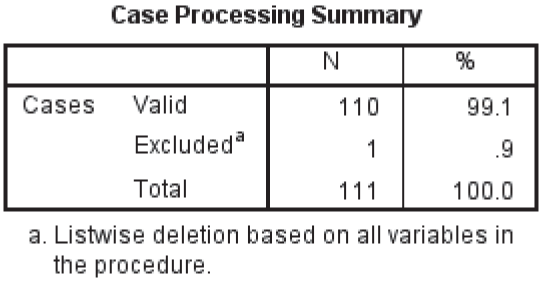

Normal P-P Plot of Obs 2 Overall Rating Domain 3
Evaluation Protocol: Differentiated Supervision
Expected Cum Prob
Observed Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0


Normal P-P Plot of Obs 2 Overall Rating Domain 3
Evaluation Protocol: Formal Observation
Expected Cum Prob
Observed Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0


Normal P-P Plot of Obs 2 Rating
Evaluation Protocol: Formal Observation
Expected Cum Prob
Observed Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0


Normal P-P Plot of Obs 2 Overall Rating Domain 3
Evaluation Protocol: Differentiated Supervision
Expected Cum Prob
Observed Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0

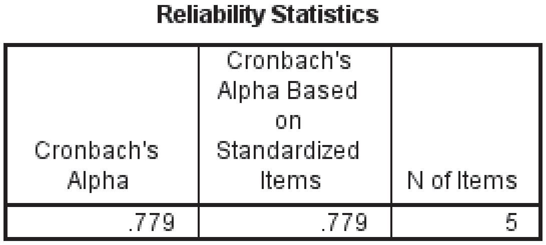
Appendix H



Case Processing Summary

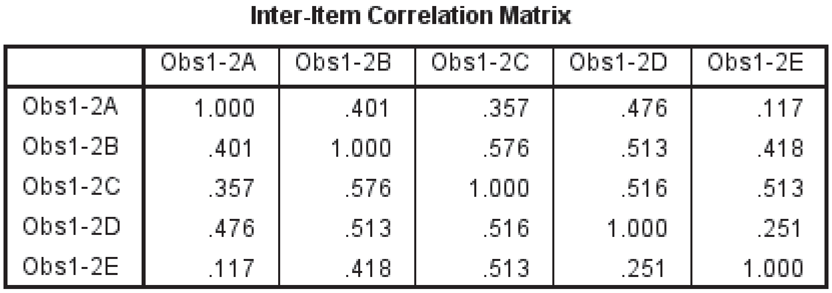
		N	%
Cases	Valid	110	99.1
	Excluded[a]	1	.9
	Total	111	100.0

a. Listwise deletion based on all variables in the procedure.



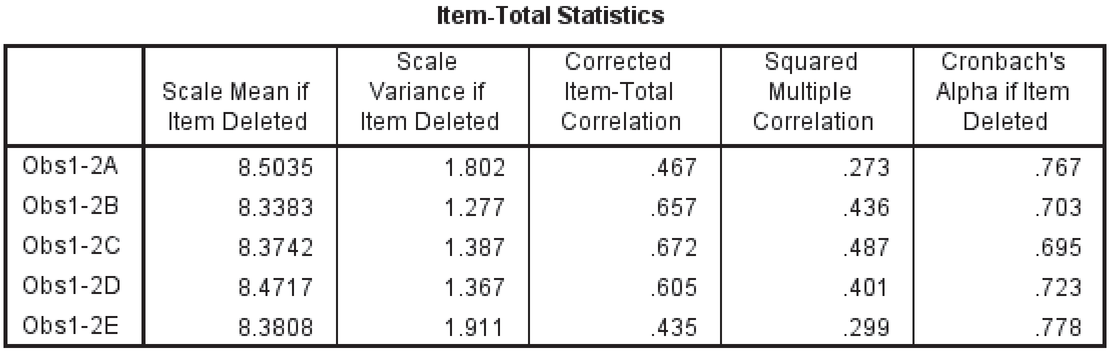
Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.779	.779	5



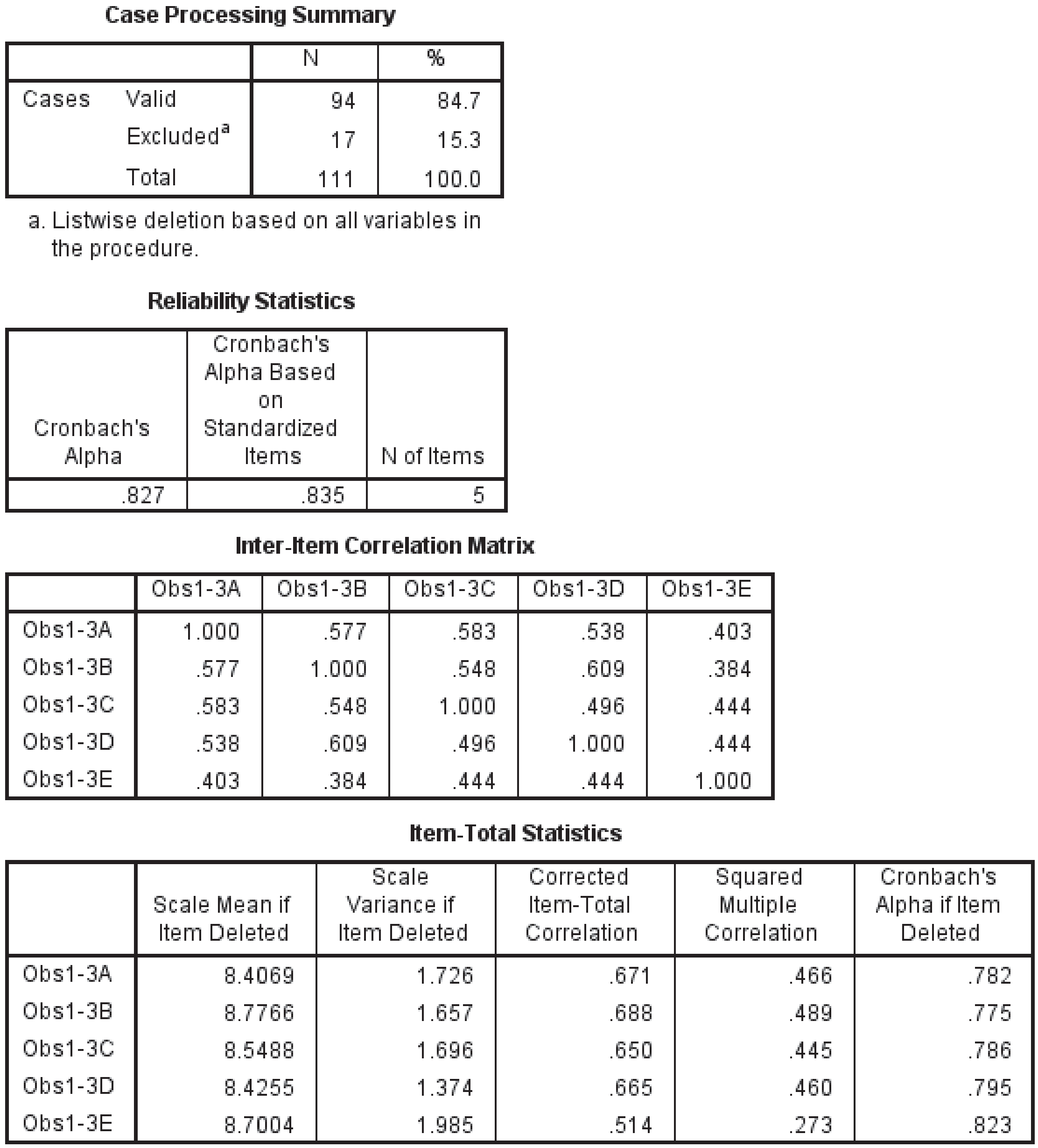
Inter-Item Correlation Matrix

	Obs1-2A	Obs1-2B	Obs1-2C	Obs1-2D	Obs1-2E
Obs1-2A	1.000	.401	.357	.476	.117
Obs1-2B	.401	1.000	.576	.513	.418
Obs1-2C	.357	.576	1.000	.516	.513
Obs1-2D	.476	.513	.516	1.000	.251
Obs1-2E	.117	.418	.513	.251	1.000



Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Obs1-2A	8.5035	1.802	.467	.273	.767
Obs1-2B	8.3383	1.277	.657	.436	.703
Obs1-2C	8.3742	1.387	.672	.487	.695
Obs1-2D	8.4717	1.367	.605	.401	.723
Obs1-2E	8.3808	1.911	.435	.299	.778



Case Processing Summary

		N	%
Cases	Valid	94	84.7
	Excluded[a]	17	15.3
	Total	111	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

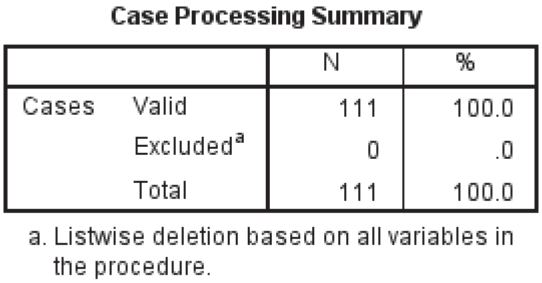
Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.827	.835	5

Inter-Item Correlation Matrix

	Obs1-3A	Obs1-3B	Obs1-3C	Obs1-3D	Obs1-3E
Obs1-3A	1.000	.577	.583	.538	.403
Obs1-3B	.577	1.000	.548	.609	.384
Obs1-3C	.583	.548	1.000	.496	.444
Obs1-3D	.538	.609	.496	1.000	.444
Obs1-3E	.403	.384	.444	.444	1.000

Item-Total Statistics

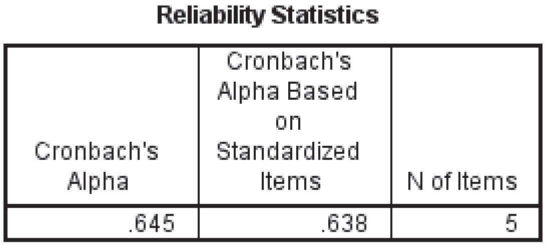
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Obs1-3A	8.4069	1.726	.671	.466	.782
Obs1-3B	8.7766	1.657	.688	.489	.775
Obs1-3C	8.5488	1.696	.650	.445	.786
Obs1-3D	8.4255	1.374	.665	.460	.795
Obs1-3E	8.7004	1.985	.514	.273	.823



Case Processing Summary

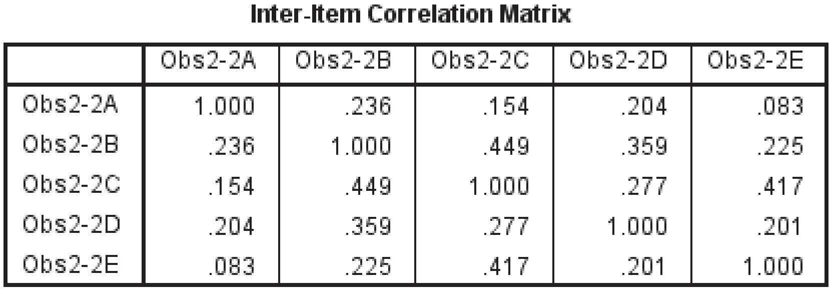
		N	%
Cases	Valid	111	100.0
	Excluded[a]	0	.0
	Total	111	100.0

a. Listwise deletion based on all variables in the procedure.



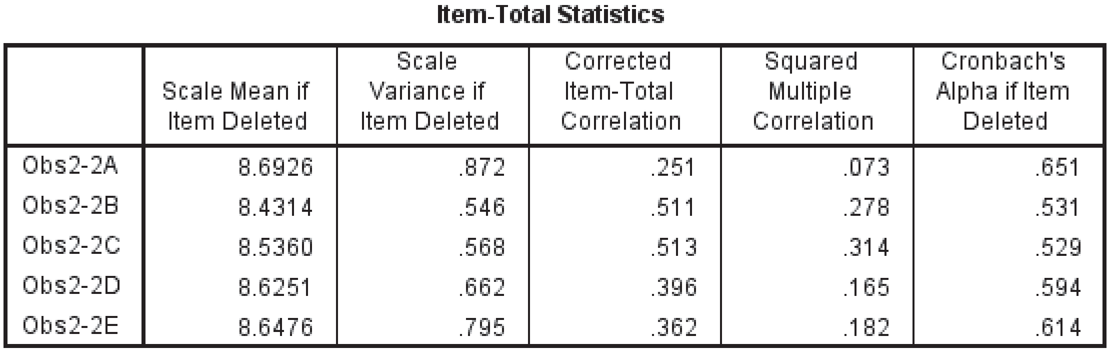
Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.645	.638	5



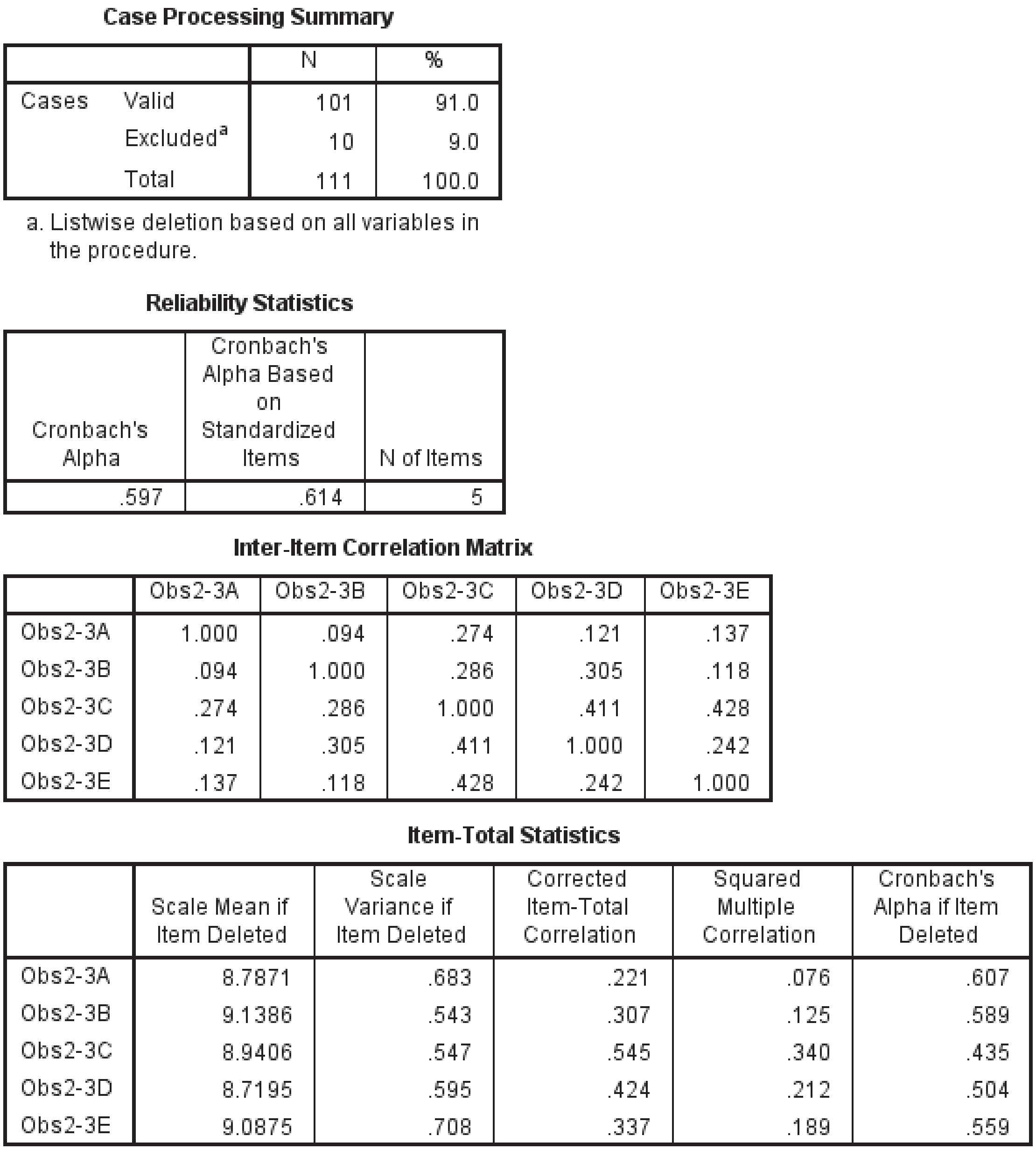
Inter-Item Correlation Matrix

	Obs2-2A	Obs2-2B	Obs2-2C	Obs2-2D	Obs2-2E
Obs2-2A	1.000	.236	.154	.204	.083
Obs2-2B	.236	1.000	.449	.359	.225
Obs2-2C	.154	.449	1.000	.277	.417
Obs2-2D	.204	.359	.277	1.000	.201
Obs2-2E	.083	.225	.417	.201	1.000



Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Obs2-2A	8.6926	.872	.251	.073	.651
Obs2-2B	8.4314	.546	.511	.278	.531
Obs2-2C	8.5360	.568	.513	.314	.529
Obs2-2D	8.6251	.662	.396	.165	.594
Obs2-2E	8.6476	.795	.362	.182	.614



Case Processing Summary

		N	%
Cases	Valid	101	91.0
	Excluded[a]	10	9.0
	Total	111	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

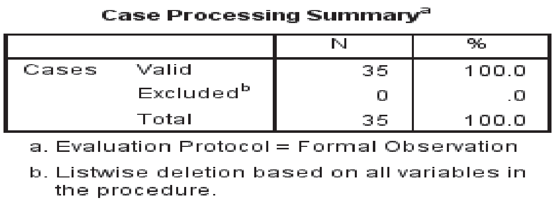
Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.597	.614	5

Inter-Item Correlation Matrix

	Obs2-3A	Obs2-3B	Obs2-3C	Obs2-3D	Obs2-3E
Obs2-3A	1.000	.094	.274	.121	.137
Obs2-3B	.094	1.000	.286	.305	.118
Obs2-3C	.274	.286	1.000	.411	.428
Obs2-3D	.121	.305	.411	1.000	.242
Obs2-3E	.137	.118	.428	.242	1.000

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Obs2-3A	8.7871	.683	.221	.076	.607
Obs2-3B	9.1386	.543	.307	.125	.589
Obs2-3C	8.9406	.547	.545	.340	.435
Obs2-3D	8.7195	.595	.424	.212	.504
Obs2-3E	9.0875	.708	.337	.189	.559

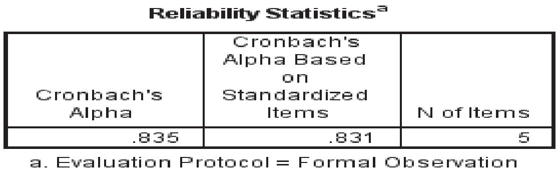


Case Processing Summary[a]

		N	%
Cases	Valid	35	100.0
	Excluded[b]	0	.0
	Total	35	100.0

a. Evaluation Protocol = Formal Observation

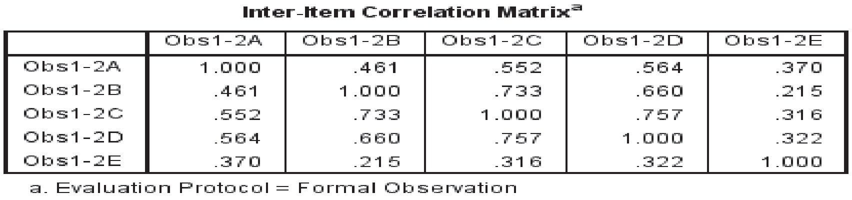
b. Listwise deletion based on all variables in the procedure.



Reliability Statistics[a]

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.835	.831	5

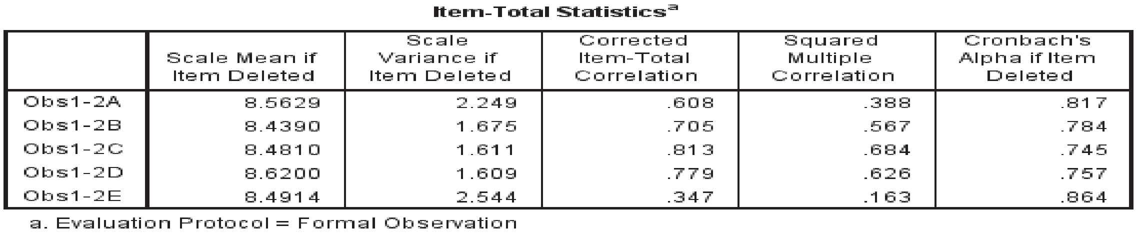
a. Evaluation Protocol = Formal Observation



Inter-Item Correlation Matrix[a]

	Obs1-2A	Obs1-2B	Obs1-2C	Obs1-2D	Obs1-2E
Obs1-2A	1.000	.461	.552	.564	.370
Obs1-2B	.461	1.000	.733	.660	.215
Obs1-2C	.552	.733	1.000	.757	.316
Obs1-2D	.564	.660	.757	1.000	.322
Obs1-2E	.370	.215	.316	.322	1.000

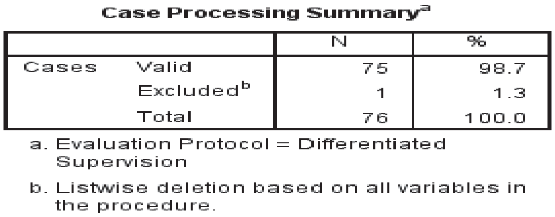
a. Evaluation Protocol = Formal Observation



Item-Total Statistics[a]

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Obs1-2A	8.5629	2.249	.608	.388	.817
Obs1-2B	8.4390	1.675	.705	.567	.784
Obs1-2C	8.4810	1.611	.813	.684	.745
Obs1-2D	8.6200	1.609	.779	.626	.757
Obs1-2E	8.4914	2.544	.347	.163	.864

a. Evaluation Protocol = Formal Observation

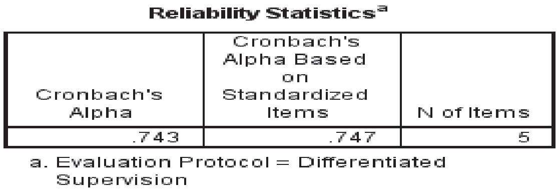


Case Processing Summary[a]

		N	%
Cases	Valid	75	98.7
	Excluded[b]	1	1.3
	Total	76	100.0

a. Evaluation Protocol = Differentiated Supervision

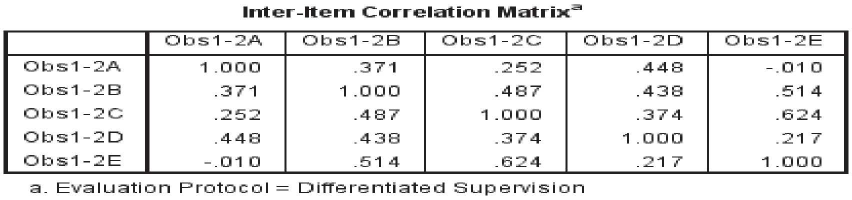
b. Listwise deletion based on all variables in the procedure.



Reliability Statistics[a]

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.743	.747	5

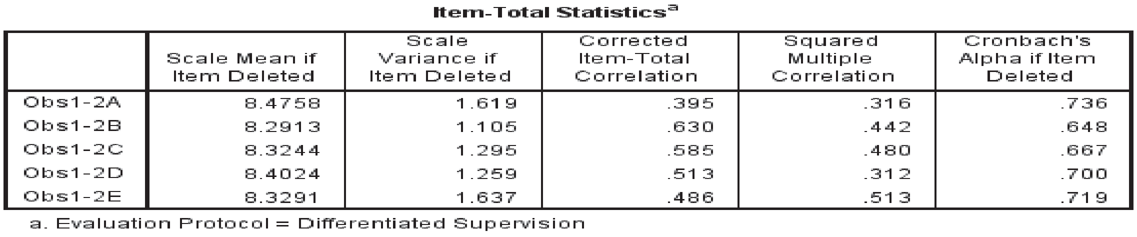
a. Evaluation Protocol = Differentiated Supervision



Inter-Item Correlation Matrix[a]

	Obs1-2A	Obs1-2B	Obs1-2C	Obs1-2D	Obs1-2E
Obs1-2A	1.000	.371	.252	.448	-.010
Obs1-2B	.371	1.000	.487	.438	.514
Obs1-2C	.252	.487	1.000	.374	.624
Obs1-2D	.448	.438	.374	1.000	.217
Obs1-2E	-.010	.514	.624	.217	1.000

a. Evaluation Protocol = Differentiated Supervision



Item-Total Statistics[a]

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Obs1-2A	8.4758	1.619	.395	.316	.736
Obs1-2B	8.2913	1.105	.630	.442	.648
Obs1-2C	8.3244	1.295	.585	.480	.667
Obs1-2D	8.4024	1.259	.513	.312	.700
Obs1-2E	8.3291	1.637	.486	.513	.719

a. Evaluation Protocol = Differentiated Supervision

Reliability Statistics[a]

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.685	.656	5

a. Evaluation Protocol = Formal Observation

Inter-Item Correlation Matrix[a]

	Obs1-3A	Obs1-3B	Obs1-3C	Obs1-3D	Obs1-3E
Obs1-3A	1.000	.446	.397	.341	-.042
Obs1-3B	.446	1.000	.464	.638	.051
Obs1-3C	.397	.464	1.000	.264	-.103
Obs1-3D	.341	.638	.264	1.000	.309
Obs1-3E	-.042	.051	-.103	.309	1.000

a. Evaluation Protocol = Formal Observation

Item-Total Statistics[a]

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Obs1-3A	8.7098	1.141	.454	.257	.630
Obs1-3B	9.0776	.983	.691	.529	.525
Obs1-3C	8.7759	1.140	.401	.273	.651
Obs1-3D	8.6695	.768	.584	.490	.575
Obs1-3E	8.9397	1.513	.106	.152	.730

a. Evaluation Protocol = Formal Observation

Case Processing Summary[a]

		N	%
Cases	Valid	65	85.5
	Excluded[b]	11	14.5
	Total	76	100.0

a. Evaluation Protocol = Differentiated Supervision

b. Listwise deletion based on all variables in the procedure.

Reliability Statistics[a]

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.859	.867	5

a. Evaluation Protocol = Differentiated Supervision

Inter-Item Correlation Matrix[a]

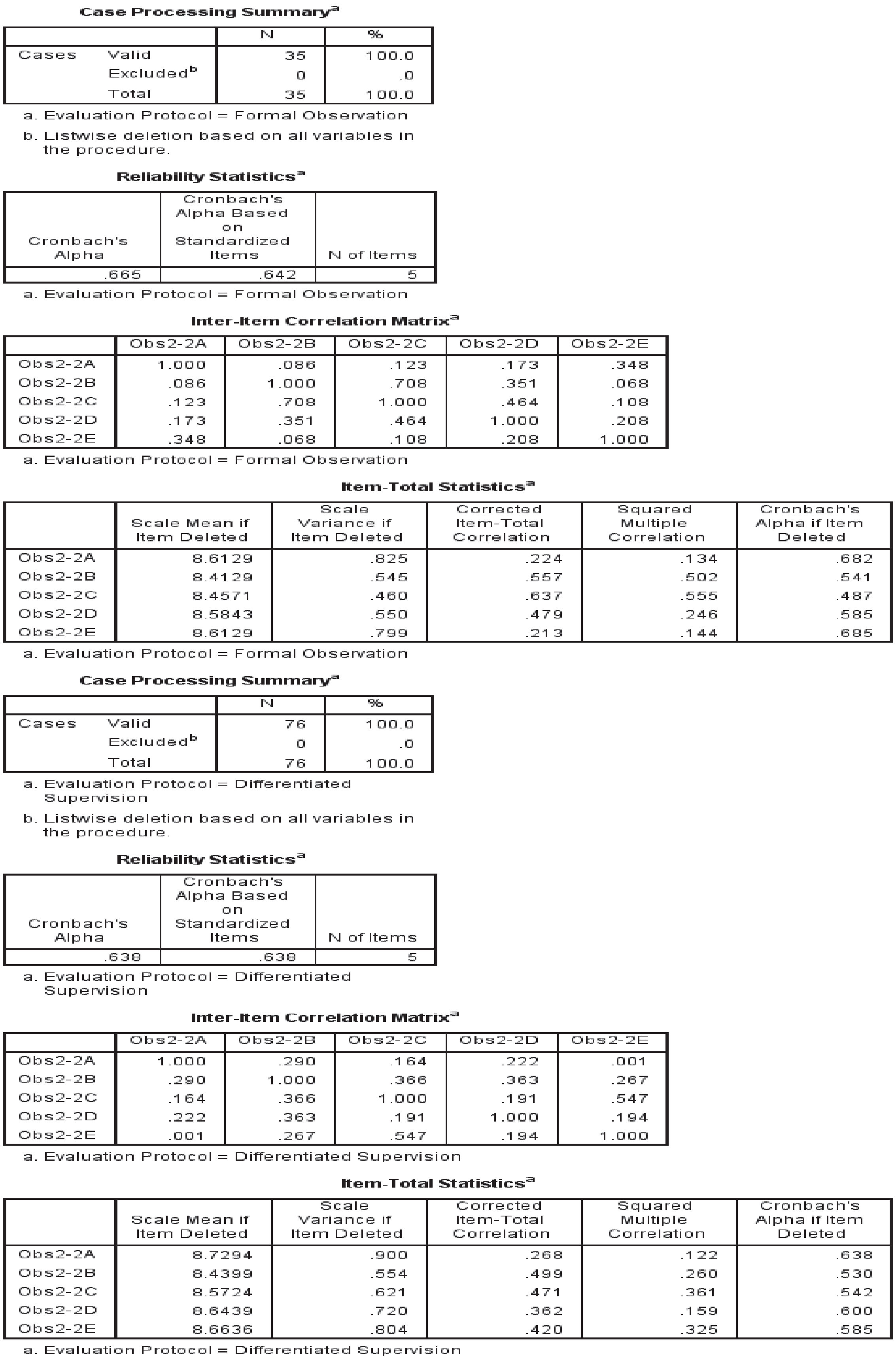
	Obs1-3A	Obs1-3B	Obs1-3C	Obs1-3D	Obs1-3E
Obs1-3A	1.000	.621	.655	.612	.514
Obs1-3B	.621	1.000	.579	.598	.461
Obs1-3C	.655	.579	1.000	.578	.563
Obs1-3D	.612	.598	.578	1.000	.477
Obs1-3E	.514	.461	.563	.477	1.000

a. Evaluation Protocol = Differentiated Supervision

Item-Total Statistics[a]

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Obs1-3A	8.2718	1.949	.742	.559	.815
Obs1-3B	8.6423	1.919	.692	.489	.826
Obs1-3C	8.4474	1.932	.724	.541	.818
Obs1-3D	8.3167	1.622	.692	.486	.838
Obs1-3E	8.5936	2.185	.596	.373	.851

a. Evaluation Protocol = Differentiated Supervision



Case Processing Summary[a]

		N	%
Cases	Valid	35	100.0
	Excluded[b]	0	.0
	Total	35	100.0

a. Evaluation Protocol = Formal Observation

b. Listwise deletion based on all variables in the procedure.

Reliability Statistics[a]

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.665	.642	5

a. Evaluation Protocol = Formal Observation

Inter-Item Correlation Matrix[a]

	Obs2-2A	Obs2-2B	Obs2-2C	Obs2-2D	Obs2-2E
Obs2-2A	1.000	.086	.123	.173	.348
Obs2-2B	.086	1.000	.708	.351	.068
Obs2-2C	.123	.708	1.000	.464	.108
Obs2-2D	.173	.351	.464	1.000	.208
Obs2-2E	.348	.068	.108	.208	1.000

a. Evaluation Protocol = Formal Observation

Item-Total Statistics[a]

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Obs2-2A	8.6129	.825	.224	.134	.682
Obs2-2B	8.4129	.545	.557	.502	.541
Obs2-2C	8.4571	.460	.637	.555	.487
Obs2-2D	8.5843	.550	.479	.246	.585
Obs2-2E	8.6129	.799	.213	.144	.685

a. Evaluation Protocol = Formal Observation

Case Processing Summary[a]

		N	%
Cases	Valid	76	100.0
	Excluded[b]	0	.0
	Total	76	100.0

a. Evaluation Protocol = Differentiated Supervision

b. Listwise deletion based on all variables in the procedure.

Reliability Statistics[a]

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.638	.638	5

a. Evaluation Protocol = Differentiated Supervision

Inter-Item Correlation Matrix[a]

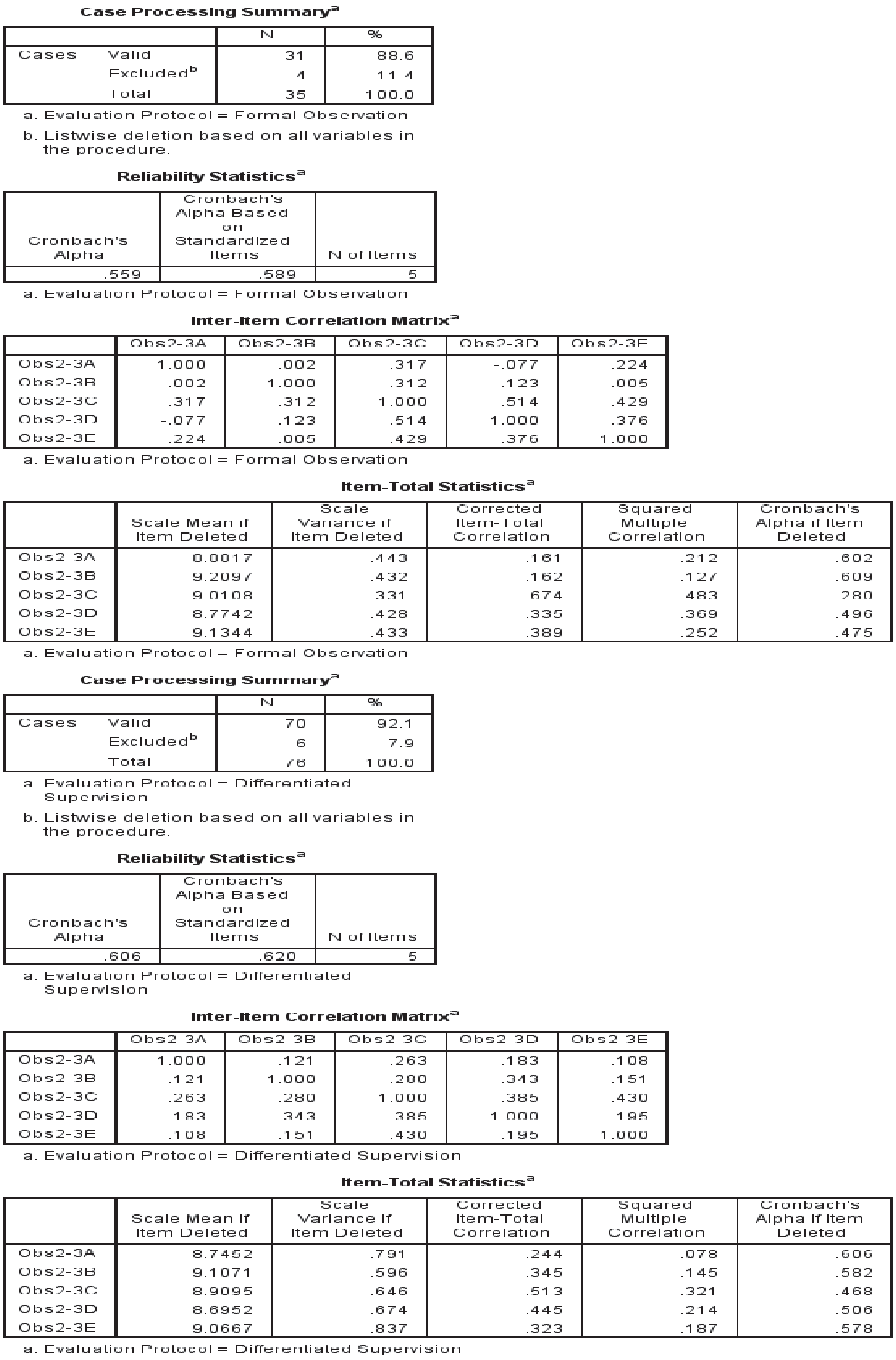
	Obs2-2A	Obs2-2B	Obs2-2C	Obs2-2D	Obs2-2E
Obs2-2A	1.000	.290	.164	.222	.001
Obs2-2B	.290	1.000	.366	.363	.267
Obs2-2C	.164	.366	1.000	.191	.547
Obs2-2D	.222	.363	.191	1.000	.194
Obs2-2E	.001	.267	.547	.194	1.000

a. Evaluation Protocol = Differentiated Supervision

Item-Total Statistics[a]

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Obs2-2A	8.7294	.900	.268	.122	.638
Obs2-2B	8.4399	.554	.499	.260	.530
Obs2-2C	8.5724	.621	.471	.361	.542
Obs2-2D	8.6439	.720	.362	.159	.600
Obs2-2E	8.6636	.804	.420	.325	.585

a. Evaluation Protocol = Differentiated Supervision



Case Processing Summary[a]

		N	%
Cases	Valid	31	88.6
	Excluded[b]	4	11.4
	Total	35	100.0

a. Evaluation Protocol = Formal Observation

b. Listwise deletion based on all variables in the procedure.

Reliability Statistics[a]

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.559	.589	5

a. Evaluation Protocol = Formal Observation

Inter-Item Correlation Matrix[a]

	Obs2-3A	Obs2-3B	Obs2-3C	Obs2-3D	Obs2-3E
Obs2-3A	1.000	.002	.317	-.077	.224
Obs2-3B	.002	1.000	.312	.123	.005
Obs2-3C	.317	.312	1.000	.514	.429
Obs2-3D	-.077	.123	.514	1.000	.376
Obs2-3E	.224	.005	.429	.376	1.000

a. Evaluation Protocol = Formal Observation

Item-Total Statistics[a]

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Obs2-3A	8.8817	.443	.161	.212	.602
Obs2-3B	9.2097	.432	.162	.127	.609
Obs2-3C	9.0108	.331	.674	.483	.280
Obs2-3D	8.7742	.428	.335	.369	.496
Obs2-3E	9.1344	.433	.389	.252	.475

a. Evaluation Protocol = Formal Observation

Case Processing Summary[a]

		N	%
Cases	Valid	70	92.1
	Excluded[b]	6	7.9
	Total	76	100.0

a. Evaluation Protocol = Differentiated Supervision

b. Listwise deletion based on all variables in the procedure.

Reliability Statistics[a]

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.606	.620	5

a. Evaluation Protocol = Differentiated Supervision

Inter-Item Correlation Matrix[a]

	Obs2-3A	Obs2-3B	Obs2-3C	Obs2-3D	Obs2-3E
Obs2-3A	1.000	.121	.263	.183	.108
Obs2-3B	.121	1.000	.280	.343	.151
Obs2-3C	.263	.280	1.000	.385	.430
Obs2-3D	.183	.343	.385	1.000	.195
Obs2-3E	.108	.151	.430	.195	1.000

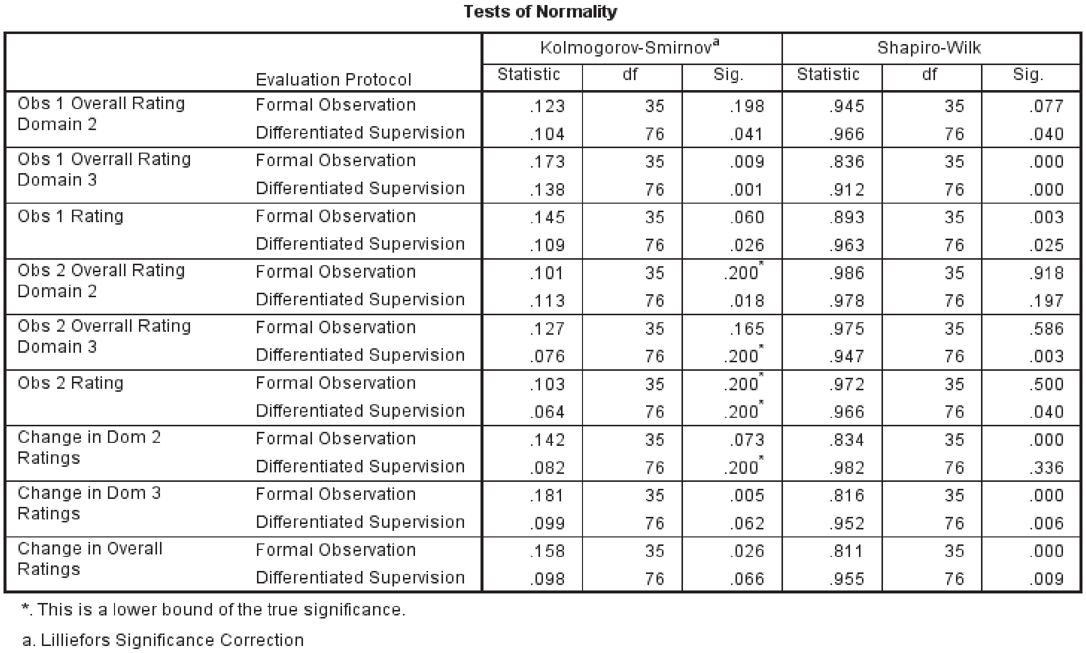
a. Evaluation Protocol = Differentiated Supervision

Item-Total Statistics[a]

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Obs2-3A	8.7452	.791	.244	.078	.606
Obs2-3B	9.1071	.596	.345	.145	.582
Obs2-3C	8.9095	.646	.513	.321	.468
Obs2-3D	8.6952	.674	.445	.214	.506
Obs2-3E	9.0667	.837	.323	.187	.578

a. Evaluation Protocol = Differentiated Supervision

Appendix I

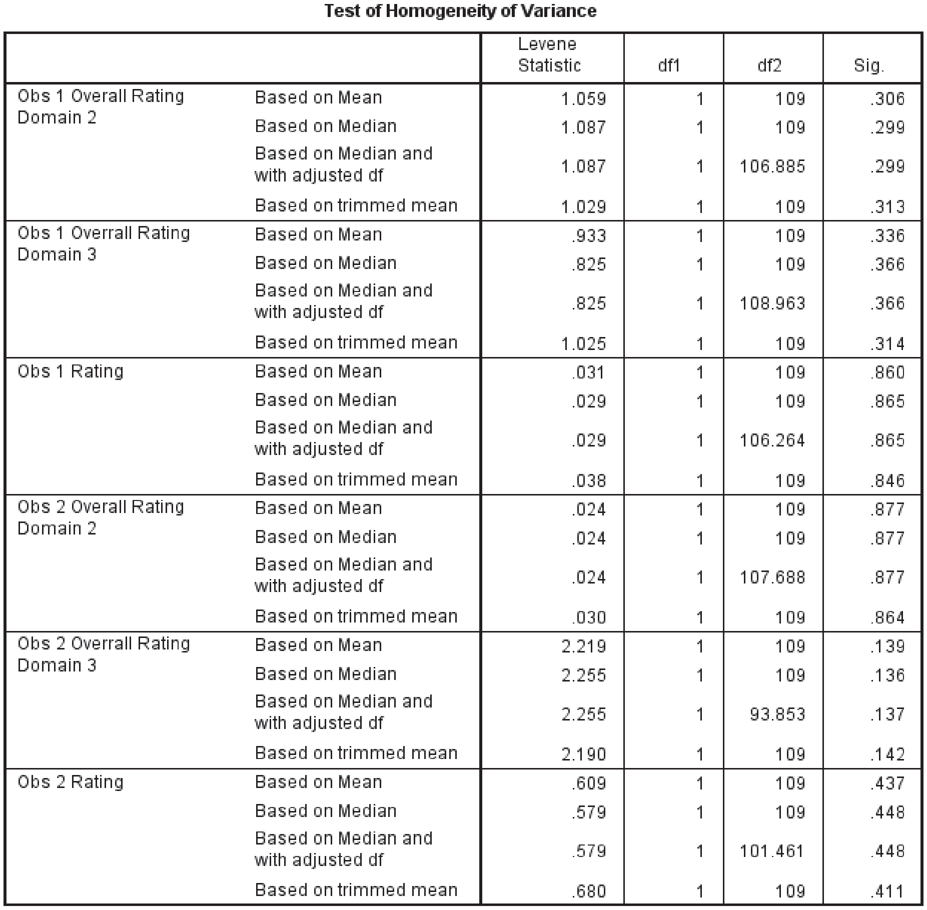


Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Evaluation Protocol	Statistic	df	Sig.	Statistic	df	Sig.
Obs 1 Overall Rating Domain 2	Formal Observation	.123	35	.198	.945	35	.077
	Differentiated Supervision	.104	76	.041	.966	76	.040
Obs 1 Overrall Rating Domain 3	Formal Observation	.173	35	.009	.836	35	.000
	Differentiated Supervision	.138	76	.001	.912	76	.000
Obs 1 Rating	Formal Observation	.145	35	.060	.893	35	.003
	Differentiated Supervision	.109	76	.026	.963	76	.025
Obs 2 Overall Rating Domain 2	Formal Observation	.101	35	.200*	.986	35	.918
	Differentiated Supervision	.113	76	.018	.978	76	.197
Obs 2 Overrall Rating Domain 3	Formal Observation	.127	35	.165	.975	35	.586
	Differentiated Supervision	.076	76	.200*	.947	76	.003
Obs 2 Rating	Formal Observation	.103	35	.200*	.972	35	.500
	Differentiated Supervision	.064	76	.200*	.966	76	.040
Change in Dom 2 Ratings	Formal Observation	.142	35	.073	.834	35	.000
	Differentiated Supervision	.082	76	.200*	.982	76	.336
Change in Dom 3 Ratings	Formal Observation	.181	35	.005	.816	35	.000
	Differentiated Supervision	.099	76	.062	.952	76	.006
Change in Overall Ratings	Formal Observation	.158	35	.026	.811	35	.000
	Differentiated Supervision	.098	76	.066	.955	76	.009

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

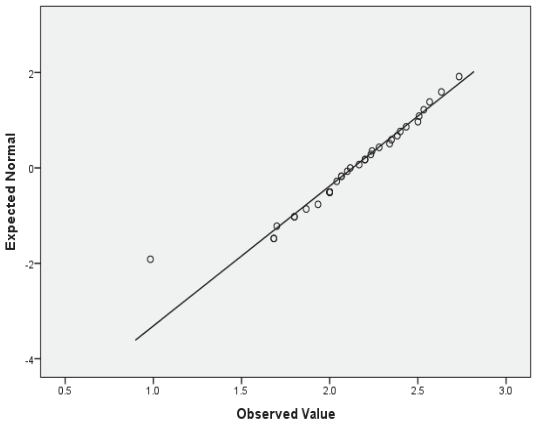


Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
Obs 1 Overall Rating Domain 2	Based on Mean	1.059	1	109	.306
	Based on Median	1.087	1	109	.299
	Based on Median and with adjusted df	1.087	1	106.885	.299
	Based on trimmed mean	1.029	1	109	.313
Obs 1 Overrall Rating Domain 3	Based on Mean	.933	1	109	.336
	Based on Median	.825	1	109	.366
	Based on Median and with adjusted df	.825	1	108.963	.366
	Based on trimmed mean	1.025	1	109	.314
Obs 1 Rating	Based on Mean	.031	1	109	.860
	Based on Median	.029	1	109	.865
	Based on Median and with adjusted df	.029	1	106.264	.865
	Based on trimmed mean	.038	1	109	.846
Obs 2 Overall Rating Domain 2	Based on Mean	.024	1	109	.877
	Based on Median	.024	1	109	.877
	Based on Median and with adjusted df	.024	1	107.688	.877
	Based on trimmed mean	.030	1	109	.864
Obs 2 Overrall Rating Domain 3	Based on Mean	2.219	1	109	.139
	Based on Median	2.255	1	109	.136
	Based on Median and with adjusted df	2.255	1	93.853	.137
	Based on trimmed mean	2.190	1	109	.142
Obs 2 Rating	Based on Mean	.609	1	109	.437
	Based on Median	.579	1	109	.448
	Based on Median and with adjusted df	.579	1	101.461	.448
	Based on trimmed mean	.680	1	109	.411

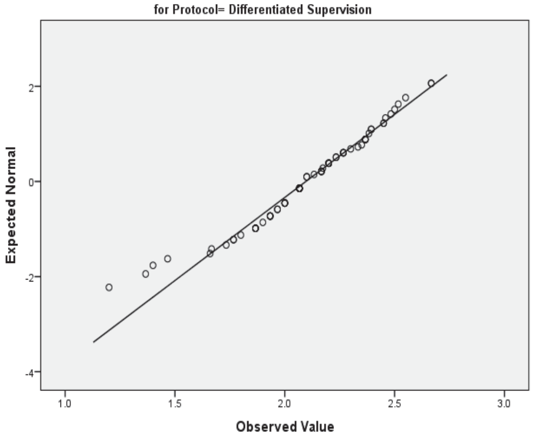
Normal Q-Q Plot of Obs 1 Overall Rating Domain 2

for Protocol= Formal Observation



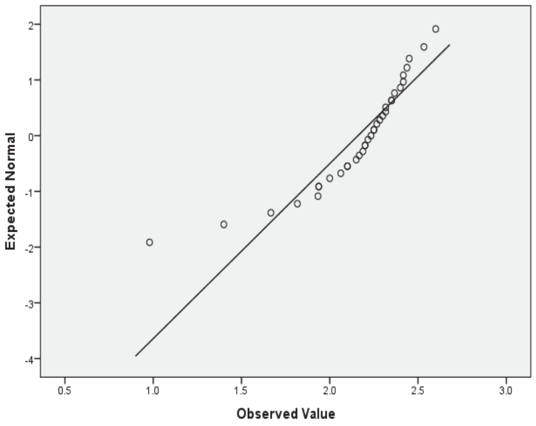
Normal Q-Q Plot of Obs 1 Overall Rating Domain 2

for Protocol= Differentiated Supervision



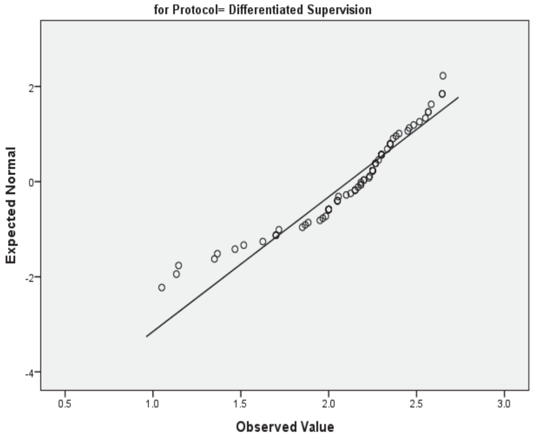
Normal Q-Q Plot of Obs 1 Overall Rating Domain 3

for Protocol= Formal Observation



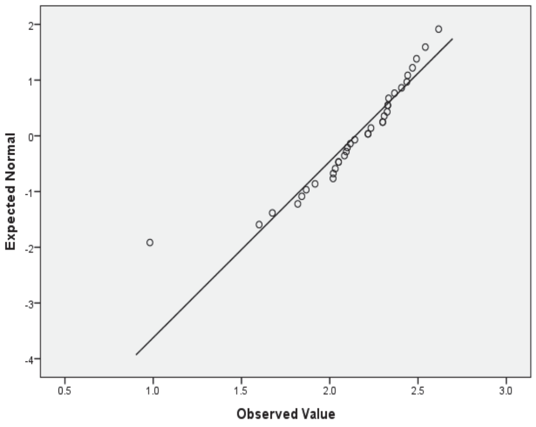
Normal Q-Q Plot of Obs 1 Overall Rating Domain 3

for Protocol= Differentiated Supervision



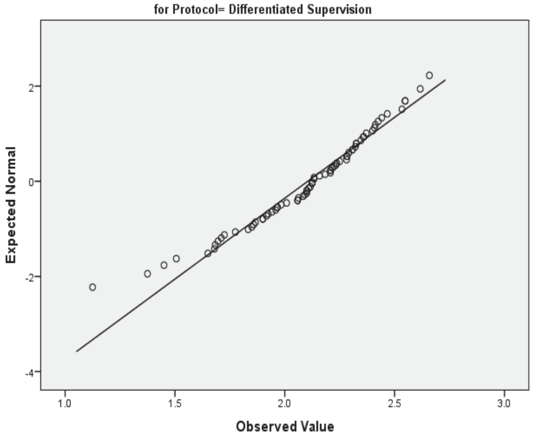
Normal Q-Q Plot of Obs 1 Rating

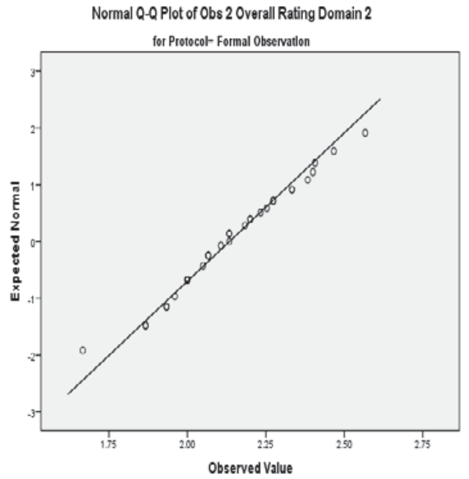
for Protocol= Formal Observation

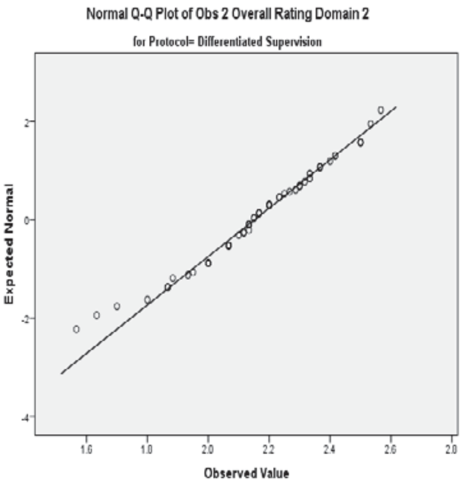


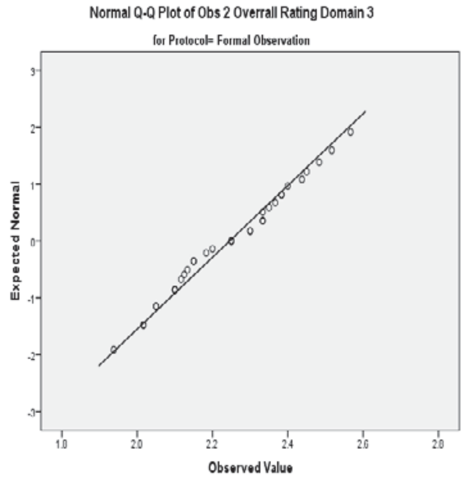
Normal Q-Q Plot of Obs 1 Rating

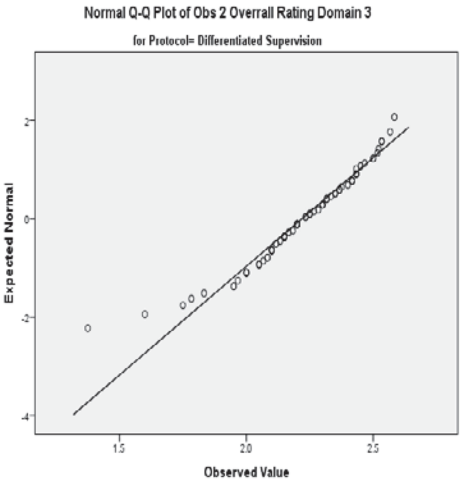
for Protocol= Differentiated Supervision

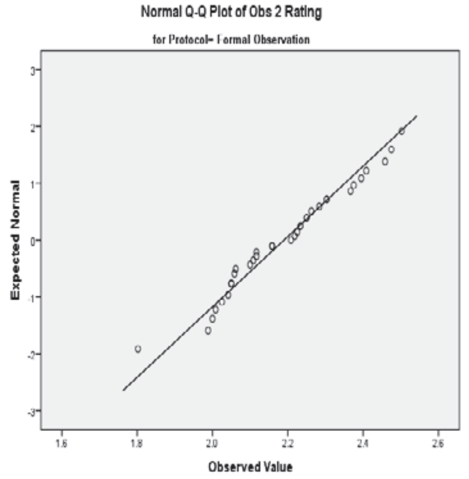


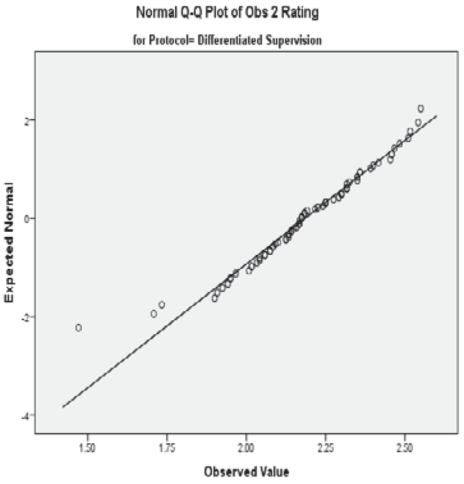

Normal Q-Q Plot of Obs 2 Overall Rating Domain 2
for Protocol= Formal Observation
Expected Normal
3
2
1
0
-1
-2
-3
1.75
2.00
2.25
2.50
2.75
Observed Value

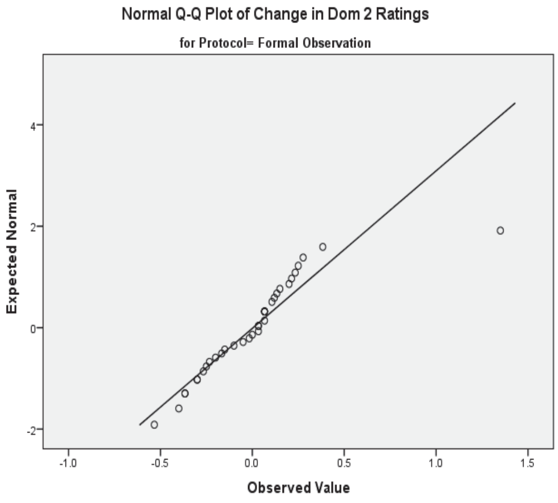

Normal Q-Q Plot of Obs 2 Overall Rating Domain 2
for Protocol= Differentiated Supervision
Expected Normal
2
0
-2
-4
1.6
1.8
2.0
2.2
2.4
2.6
2.8
Observed Value


Normal Q-Q Plot of Obs 2 Overall Rating Domain 3
for Protocol= Formal Observation
Expected Normal
3
2
1
0
-1
-2
-3
1.8
2.0
2.2
2.4
2.6
2.8
Observed Value

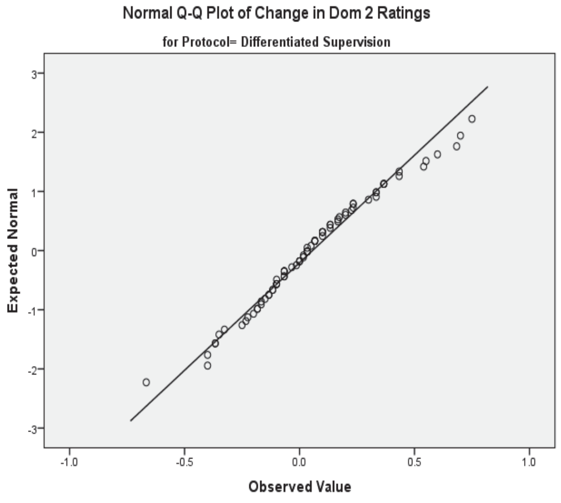

Normal Q-Q Plot of Obs 2 Overall Rating Domain 3
for Protocol= Differentiated Supervision
Expected Normal
2
0
-2
-4
1.5
2.0
2.5
Observed Value


Normal Q-Q Plot of Obs 2 Rating
for Protocol= Formal Observation
Expected Normal
3
2
1
0
-1
-2
-3
1.6
1.8
2.0
2.2
2.4
2.6
Observed Value

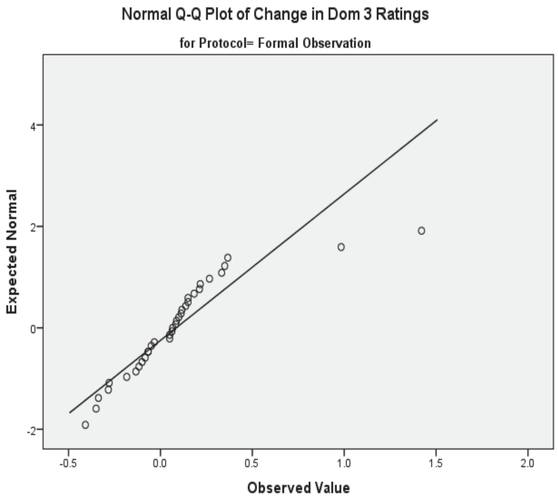

Normal Q-Q Plot of Obs 2 Rating
for Protocol= Differentiated Supervision
Expected Normal
2
0
-2
-4
1.50
1.75
2.00
2.25
2.50
Observed Value


Normal Q-Q Plot of Change in Dom 2 Ratings

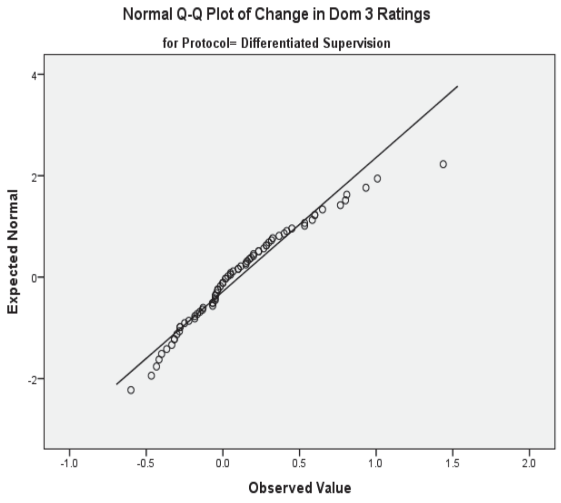
for Protocol= Formal Observation


Normal Q-Q Plot of Change in Dom 2 Ratings

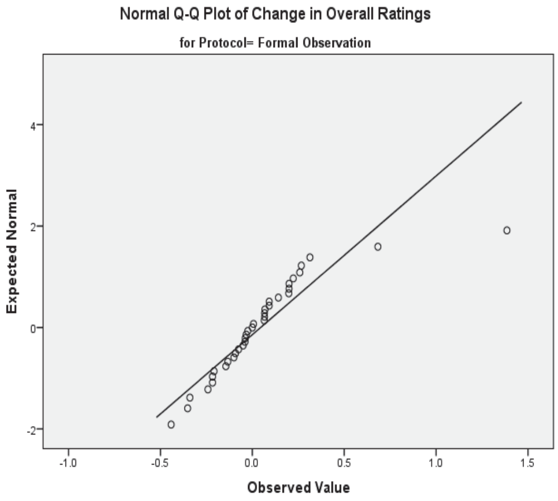
for Protocol= Differentiated Supervision


Normal Q-Q Plot of Change in Dom 3 Ratings

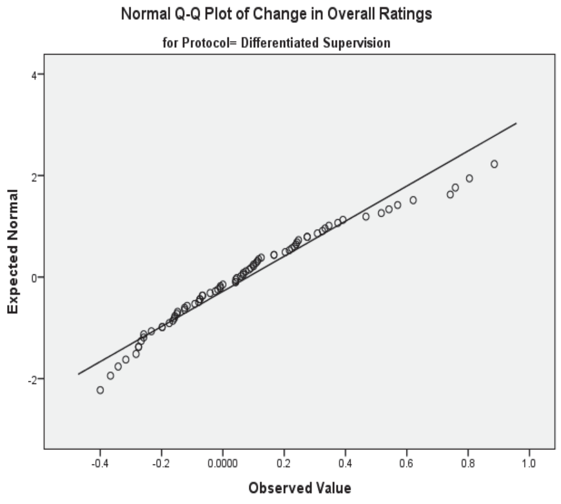
for Protocol= Formal Observation


Normal Q-Q Plot of Change in Dom 3 Ratings

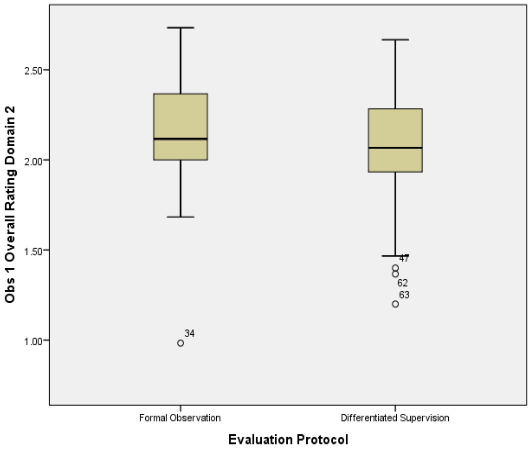
for Protocol= Differentiated Supervision

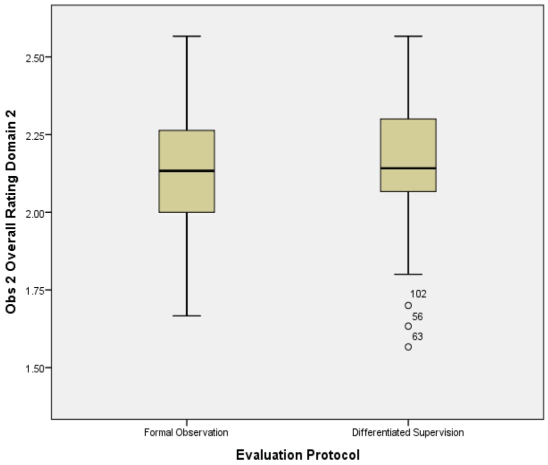

Normal Q-Q Plot of Change in Overall Ratings

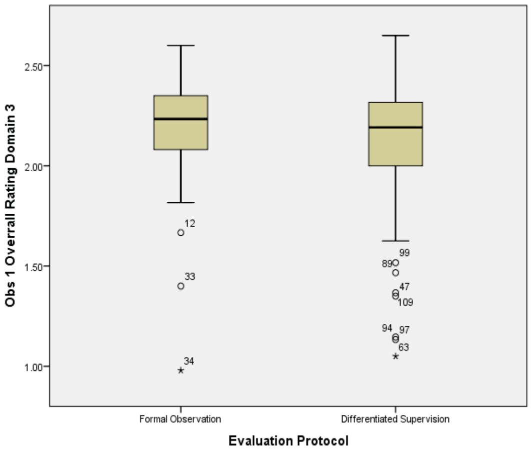
for Protocol= Formal Observation

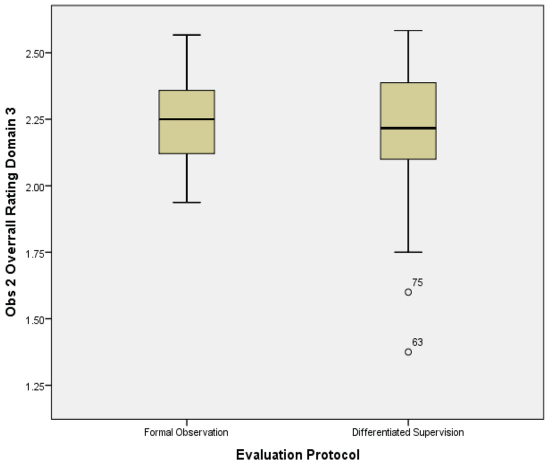

Normal Q-Q Plot of Change in Overall Ratings

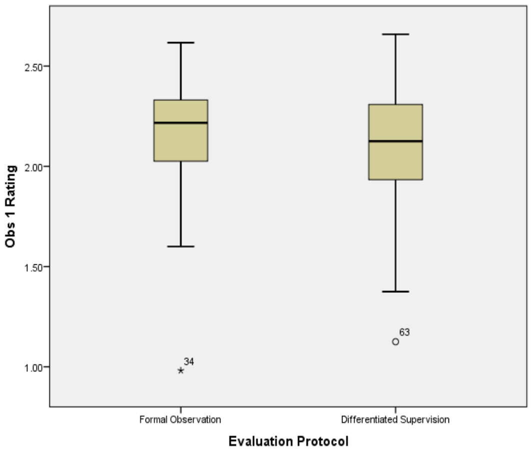
for Protocol= Differentiated Supervision

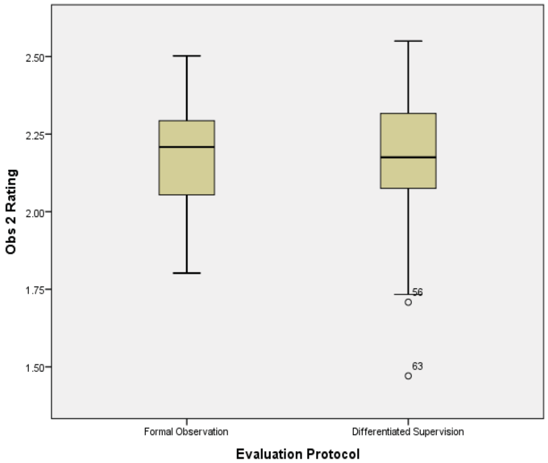

Obs 1 Overall Rating Domain 2
2.50
2.00
1.50
1.00
34
47
62
63
Formal Observation
Differentiated Supervision
Evaluation Protocol

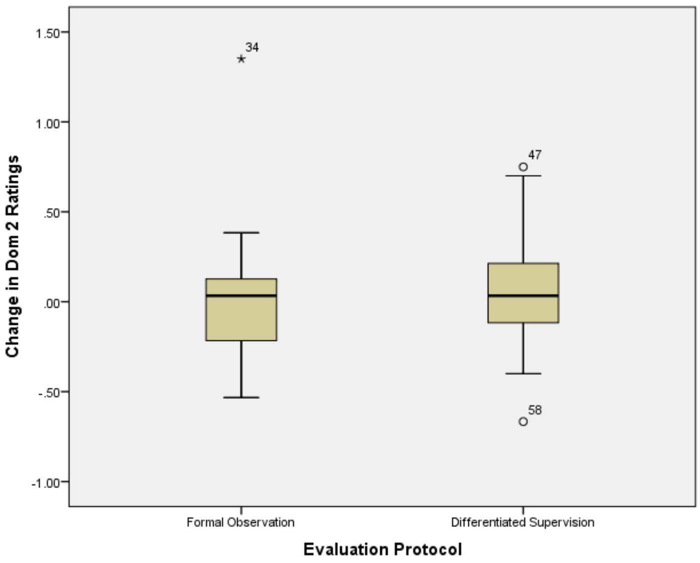

Obs 2 Overall Rating Domain 2
2.50
2.25
2.00
1.75
1.50
102
56
63
Formal Observation
Differentiated Supervision
Evaluation Protocol

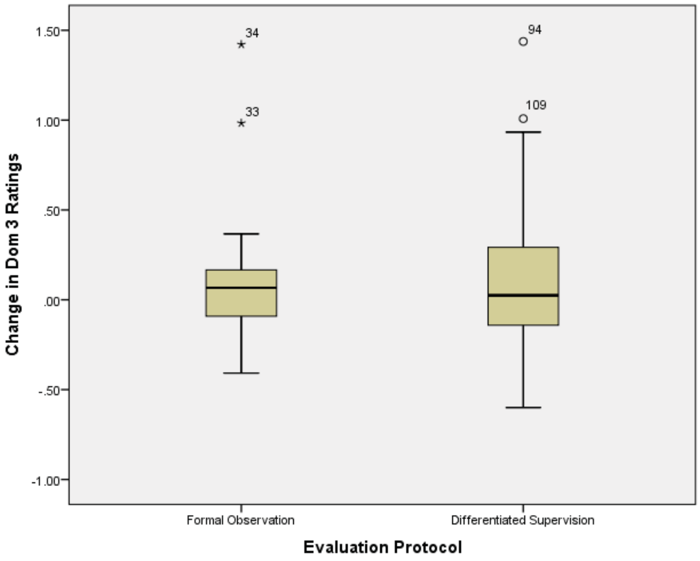

Obs 1 Overall Rating Domain 3
2.50
2.00
1.50
1.00
12
33
34
99
89
47
109
94
97
63
Formal Observation
Differentiated Supervision
Evaluation Protocol

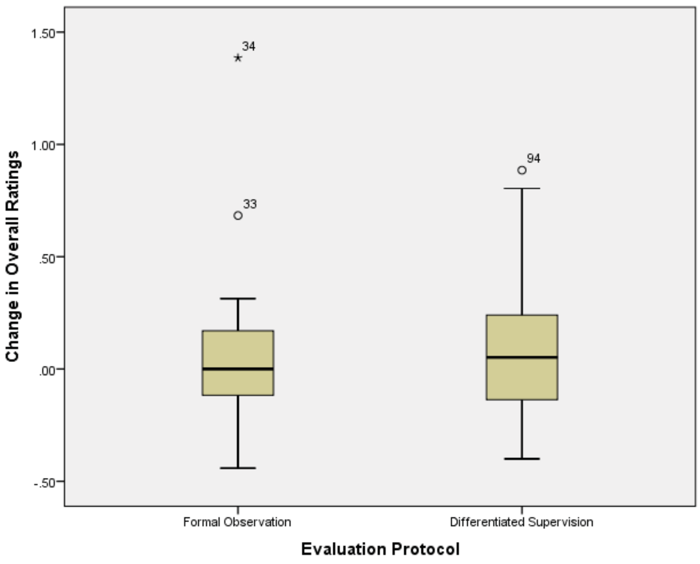

Obs 2 Overall Rating Domain 3
2.50
2.25
2.00
1.75
1.50
1.25
75
63
Formal Observation
Differentiated Supervision
Evaluation Protocol

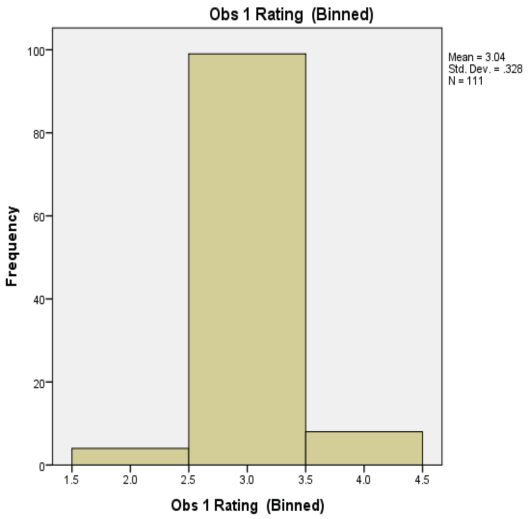

Obs 1 Rating
2.50
2.00
1.50
1.00
34
63
Formal Observation
Differentiated Supervision
Evaluation Protocol

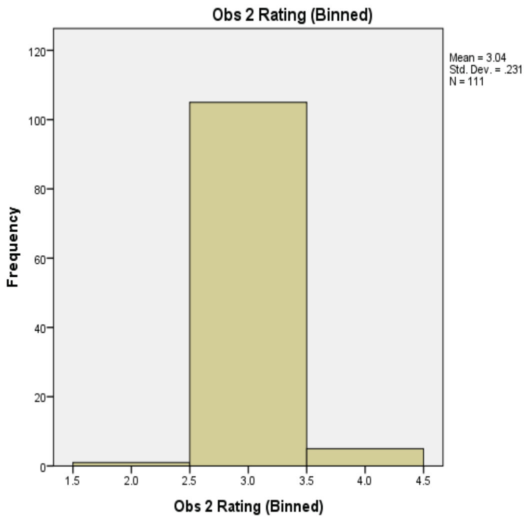

Obs 2 Rating
2.50
2.25
2.00
1.75
1.50
56
63
Formal Observation
Differentiated Supervision
Evaluation Protocol

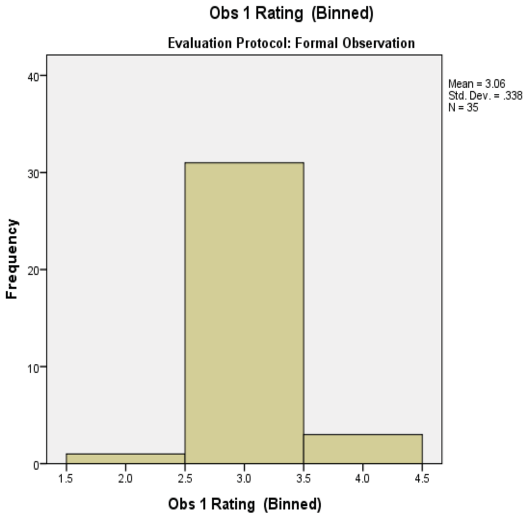

1.50
1.00
.50
.00
-.50
-1.00
Change in Dom 2 Ratings
34
47
58
Formal Observation
Differentiated Supervision
Evaluation Protocol

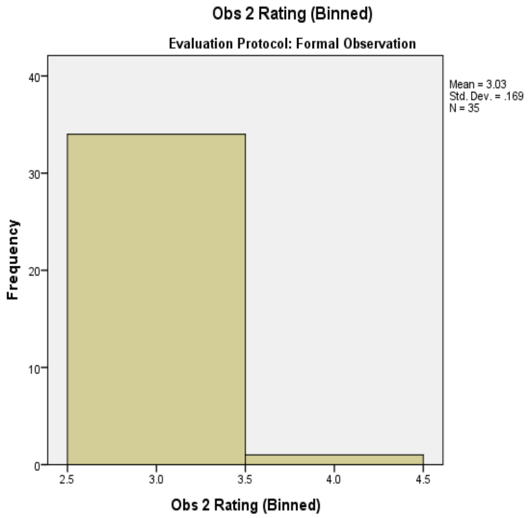

1.50
1.00
.50
.00
-.50
-1.00
Change in Dom 3 Ratings
34
33
94
109
Formal Observation
Differentiated Supervision
Evaluation Protocol

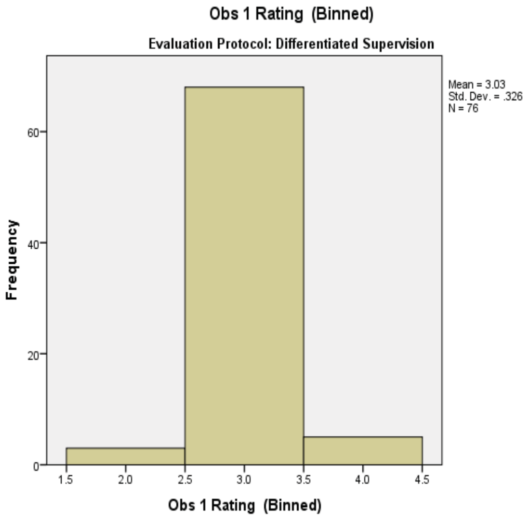

1.50
1.00
.50
.00
-.50
Change in Overall Ratings
34
33
94
Formal Observation
Differentiated Supervision
Evaluation Protocol

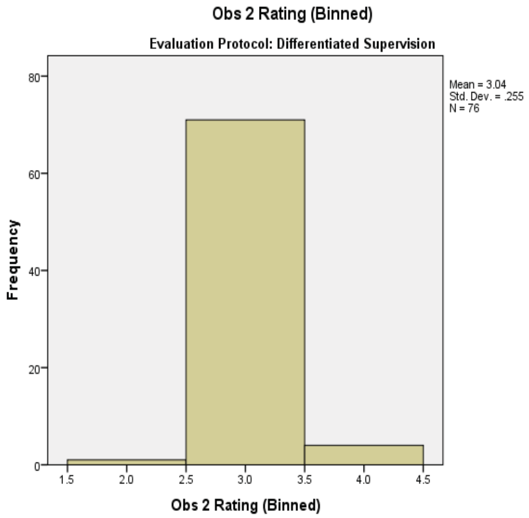

Obs 1 Rating (Binned)
Mean = 3.04
Std. Dev. = .328
N = 111
Frequency
0
20
40
60
80
100
1.5
2.0
2.5
3.0
3.5
4.0
4.5
Obs 1 Rating (Binned)

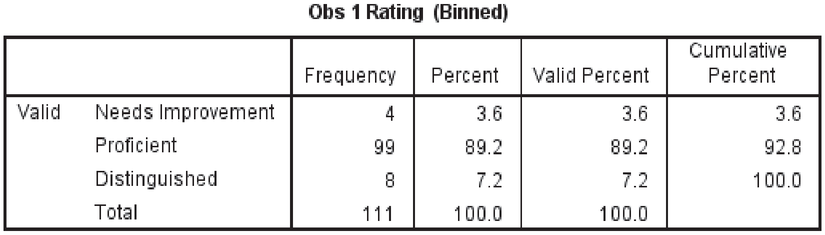

Obs 2 Rating (Binned)
Mean = 3.04
Std. Dev. = .231
N = 111
Frequency
0
20
40
60
80
100
120
1.5
2.0
2.5
3.0
3.5
4.0
4.5
Obs 2 Rating (Binned)


Obs 1 Rating (Binned)
Evaluation Protocol: Formal Observation
Mean = 3.06
Std. Dev. = .338
N = 35
Frequency
0
10
20
30
40
1.5
2.0
2.5
3.0
3.5
4.0
4.5
Obs 1 Rating (Binned)


Obs 2 Rating (Binned)
Evaluation Protocol: Formal Observation
Mean = 3.03
Std. Dev. = .169
N = 35
Frequency
0
10
20
30
40
2.5
3.0
3.5
4.0
4.5
Obs 2 Rating (Binned)

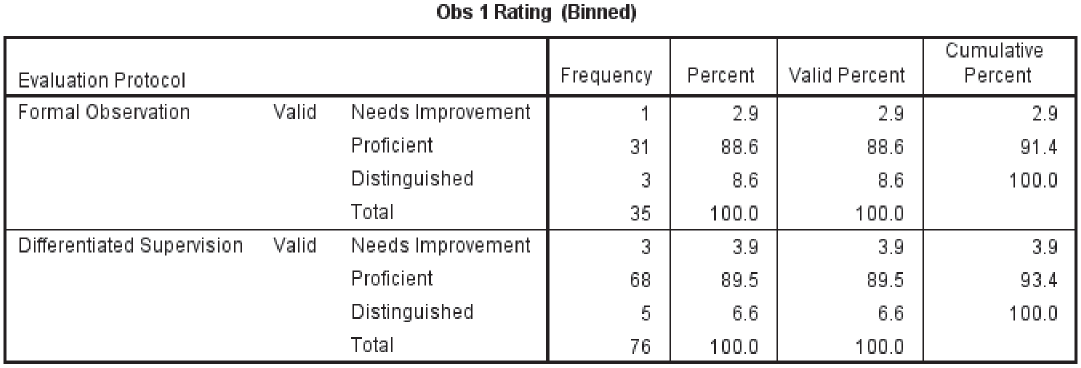

Obs 1 Rating (Binned)
Evaluation Protocol: Differentiated Supervision
Mean = 3.03
Std. Dev. = .326
N = 76
Frequency
0
20
40
60
1.5
2.0
2.5
3.0
3.5
4.0
4.5
Obs 1 Rating (Binned)


Obs 2 Rating (Binned)
Evaluation Protocol: Differentiated Supervision
Mean = 3.04
Std. Dev. = .255
N = 76
Frequency
0
20
40
60
80
1.5
2.0
2.5
3.0
3.5
4.0
4.5
Obs 2 Rating (Binned)



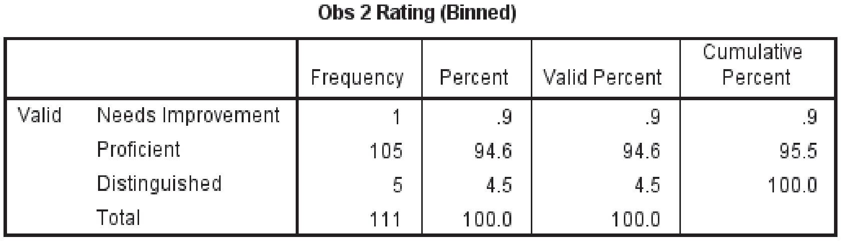
Obs 1 Rating (Binned)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Needs Improvement	4	3.6	3.6	3.6
	Proficient	99	89.2	89.2	92.8
	Distinguished	8	7.2	7.2	100.0
	Total	111	100.0	100.0	



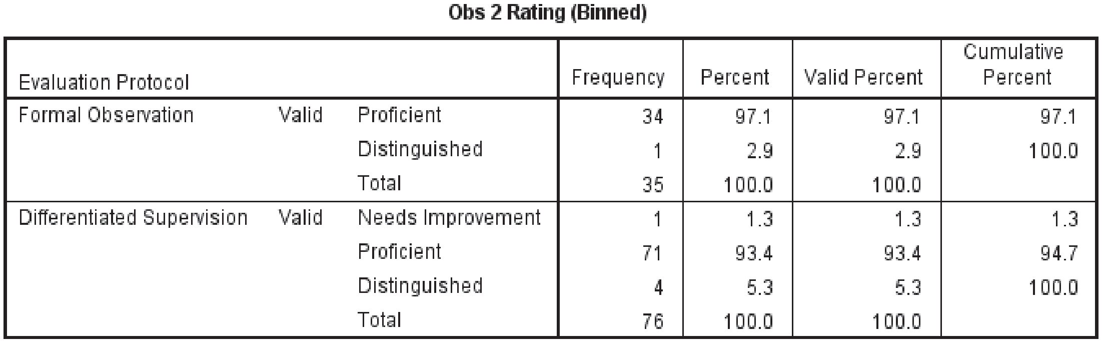
Obs 1 Rating (Binned)

Evaluation Protocol			Frequency	Percent	Valid Percent	Cumulative Percent
Formal Observation	Valid	Needs Improvement	1	2.9	2.9	2.9
		Proficient	31	88.6	88.6	91.4
		Distinguished	3	8.6	8.6	100.0
		Total	35	100.0	100.0	
Differentiated Supervision	Valid	Needs Improvement	3	3.9	3.9	3.9
		Proficient	68	89.5	89.5	93.4
		Distinguished	5	6.6	6.6	100.0
		Total	76	100.0	100.0	



Obs 2 Rating (Binned)

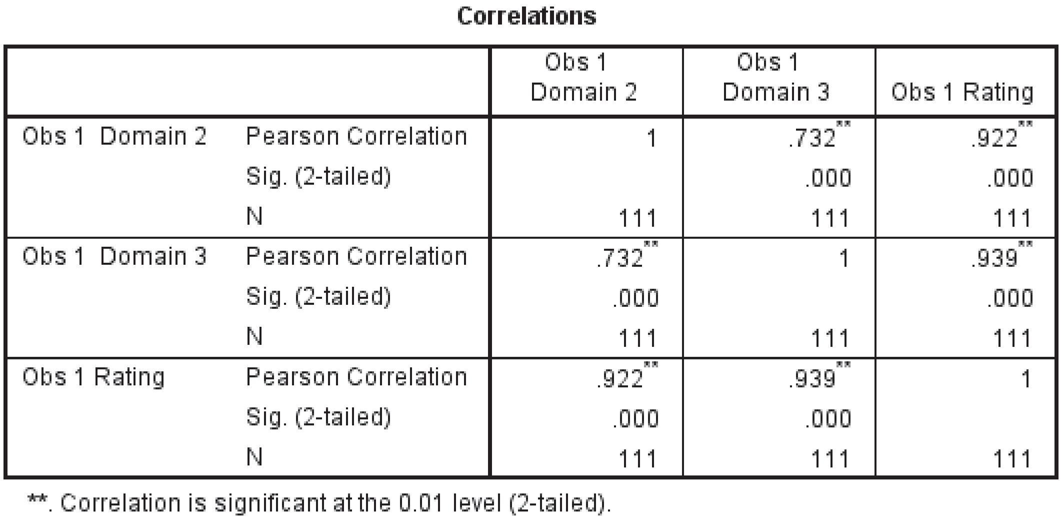
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Needs Improvement	1	.9	.9	.9
	Proficient	105	94.6	94.6	95.5
	Distinguished	5	4.5	4.5	100.0
	Total	111	100.0	100.0	



Obs 2 Rating (Binned)

Evaluation Protocol			Frequency	Percent	Valid Percent	Cumulative Percent
Formal Observation	Valid	Proficient	34	97.1	97.1	97.1
		Distinguished	1	2.9	2.9	100.0
		Total	35	100.0	100.0	
Differentiated Supervision	Valid	Needs Improvement	1	1.3	1.3	1.3
		Proficient	71	93.4	93.4	94.7
		Distinguished	4	5.3	5.3	100.0
		Total	76	100.0	100.0	

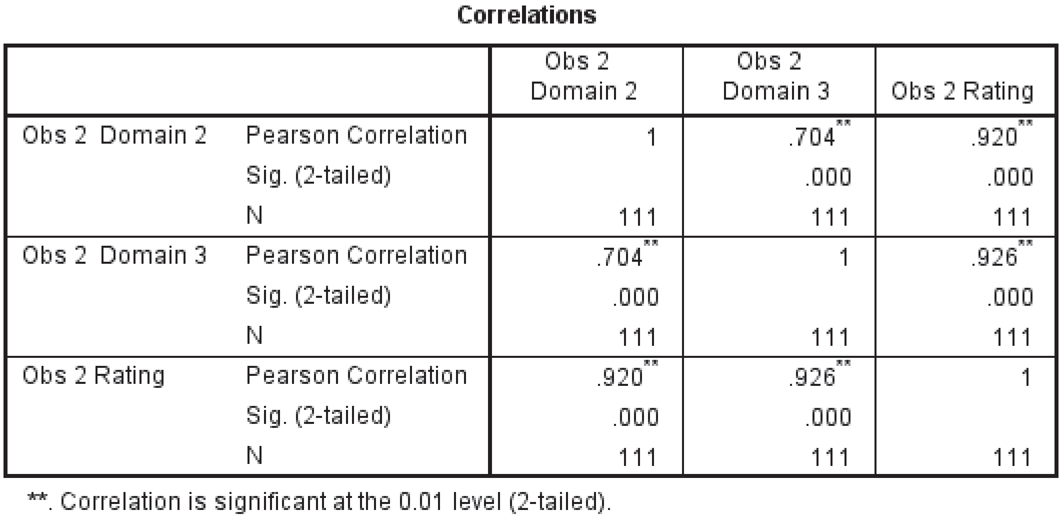
Appendix J



Correlations

		Obs 1 Domain 2	Obs 1 Domain 3	Obs 1 Rating
Obs 1 Domain 2	Pearson Correlation	1	.732**	.922**
	Sig. (2-tailed)		.000	.000
	N	111	111	111
Obs 1 Domain 3	Pearson Correlation	.732**	1	.939**
	Sig. (2-tailed)	.000		.000
	N	111	111	111
Obs 1 Rating	Pearson Correlation	.922**	.939**	1
	Sig. (2-tailed)	.000	.000	
	N	111	111	111

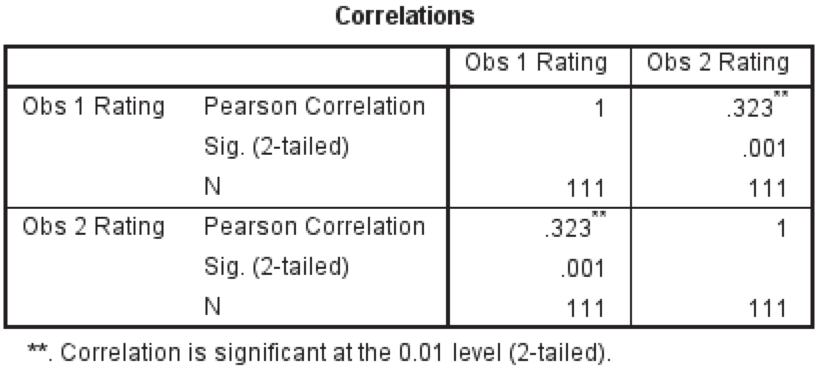
**. Correlation is significant at the 0.01 level (2-tailed).



Correlations

		Obs 2 Domain 2	Obs 2 Domain 3	Obs 2 Rating
Obs 2 Domain 2	Pearson Correlation	1	.704**	.920**
	Sig. (2-tailed)		.000	.000
	N	111	111	111
Obs 2 Domain 3	Pearson Correlation	.704**	1	.926**
	Sig. (2-tailed)	.000		.000
	N	111	111	111
Obs 2 Rating	Pearson Correlation	.920**	.926**	1
	Sig. (2-tailed)	.000	.000	
	N	111	111	111

**. Correlation is significant at the 0.01 level (2-tailed).

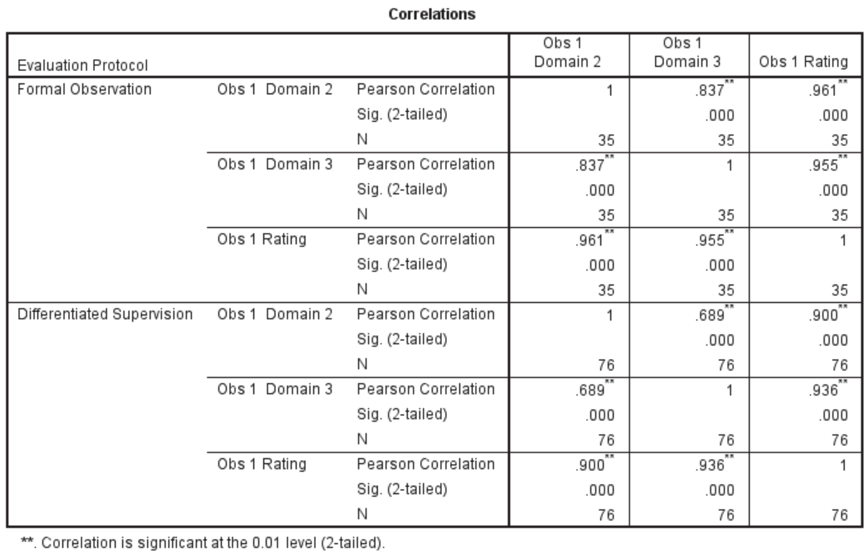


Correlations

		Obs 1 Rating	Obs 2 Rating
Obs 1 Rating	Pearson Correlation	1	.323**
	Sig. (2-tailed)		.001
	N	111	111
Obs 2 Rating	Pearson Correlation	.323**	1
	Sig. (2-tailed)	.001	
	N	111	111

**. Correlation is significant at the 0.01 level (2-tailed).

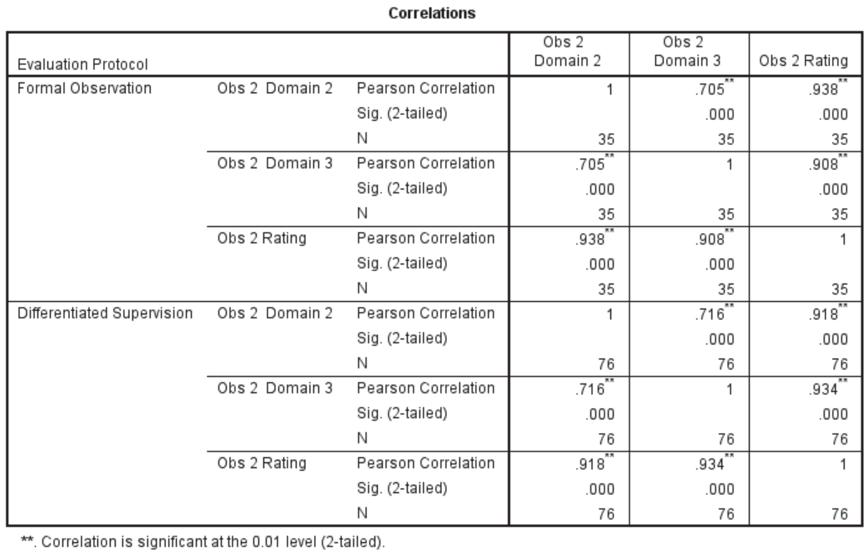
Appendix K



Correlations

Evaluation Protocol			Obs 1 Domain 2	Obs 1 Domain 3	Obs 1 Rating
Formal Observation	Obs 1 Domain 2	Pearson Correlation	1	.837**	.961**
		Sig. (2-tailed)		.000	.000
		N	35	35	35
	Obs 1 Domain 3	Pearson Correlation	.837**	1	.955**
		Sig. (2-tailed)	.000		.000
		N	35	35	35
	Obs 1 Rating	Pearson Correlation	.961**	.955**	1
		Sig. (2-tailed)	.000	.000	
		N	35	35	35
Differentiated Supervision	Obs 1 Domain 2	Pearson Correlation	1	.689**	.900**
		Sig. (2-tailed)		.000	.000
		N	76	76	76
	Obs 1 Domain 3	Pearson Correlation	.689**	1	.936**
		Sig. (2-tailed)	.000		.000
		N	76	76	76
	Obs 1 Rating	Pearson Correlation	.900**	.936**	1
		Sig. (2-tailed)	.000	.000	
		N	76	76	76

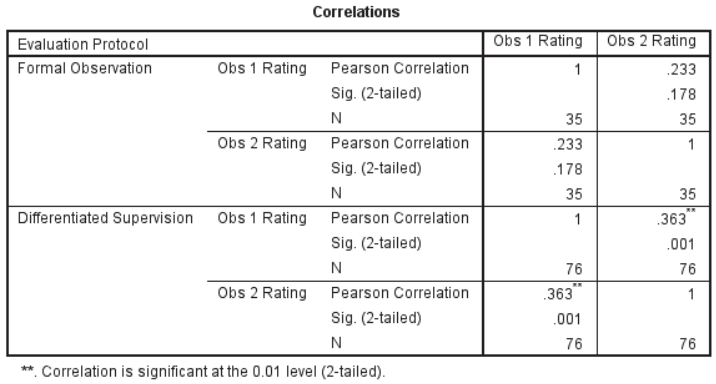
**. Correlation is significant at the 0.01 level (2-tailed).



Correlations

Evaluation Protocol			Obs 2 Domain 2	Obs 2 Domain 3	Obs 2 Rating
Formal Observation	Obs 2 Domain 2	Pearson Correlation	1	.705**	.938**
		Sig. (2-tailed)		.000	.000
		N	35	35	35
	Obs 2 Domain 3	Pearson Correlation	.705**	1	.908**
		Sig. (2-tailed)	.000		.000
		N	35	35	35
	Obs 2 Rating	Pearson Correlation	.938**	.908**	1
		Sig. (2-tailed)	.000	.000	
		N	35	35	35
Differentiated Supervision	Obs 2 Domain 2	Pearson Correlation	1	.716**	.918**
		Sig. (2-tailed)		.000	.000
		N	76	76	76
	Obs 2 Domain 3	Pearson Correlation	.716**	1	.934**
		Sig. (2-tailed)	.000		.000
		N	76	76	76
	Obs 2 Rating	Pearson Correlation	.918**	.934**	1
		Sig. (2-tailed)	.000	.000	
		N	76	76	76

**. Correlation is significant at the 0.01 level (2-tailed).



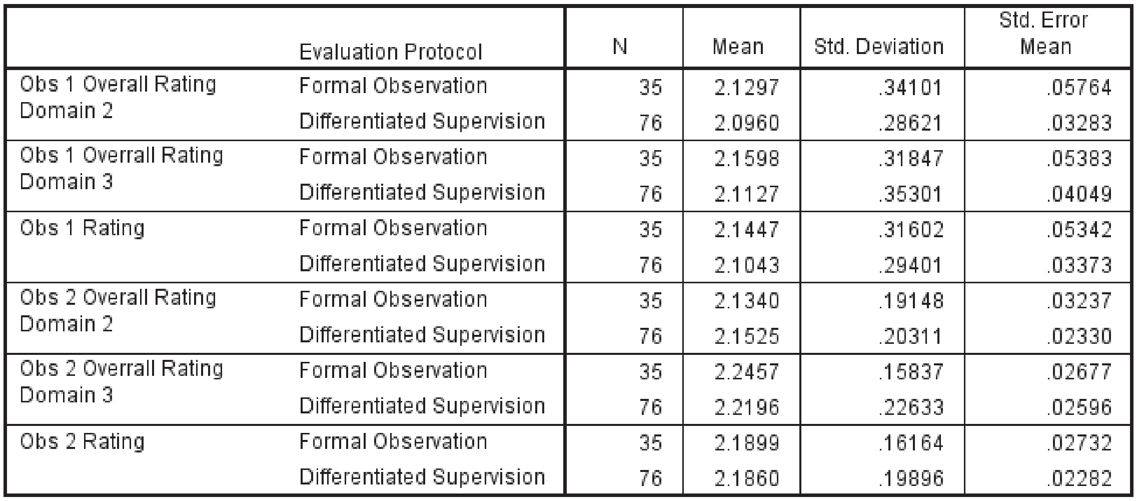
Correlations

Evaluation Protocol			Obs 1 Rating	Obs 2 Rating
Formal Observation	Obs 1 Rating	Pearson Correlation	1	.233
		Sig. (2-tailed)		.178
		N	35	35
	Obs 2 Rating	Pearson Correlation	.233	1
		Sig. (2-tailed)	.178	
		N	35	35
Differentiated Supervision	Obs 1 Rating	Pearson Correlation	1	.363**
		Sig. (2-tailed)		.001
		N	76	76
	Obs 2 Rating	Pearson Correlation	.363**	1
		Sig. (2-tailed)	.001	
		N	76	76

**. Correlation is significant at the 0.01 level (2-tailed).

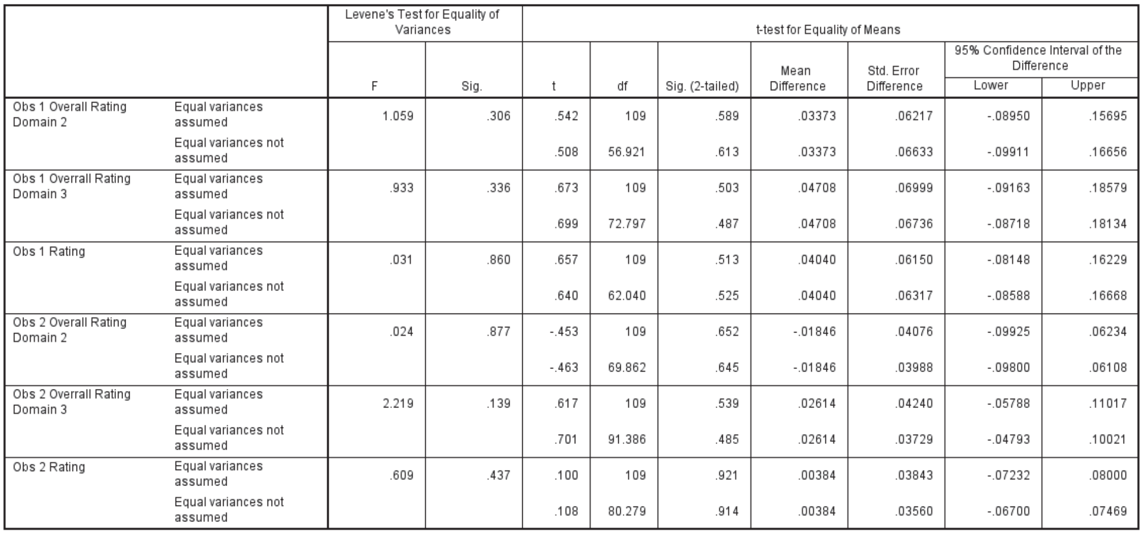
Appendix L

Group Statistics



	Evaluation Protocol	N	Mean	Std. Deviation	Std. Error Mean
Obs 1 Overall Rating Domain 2	Formal Observation	35	2.1297	.34101	.05764
	Differentiated Supervision	76	2.0960	.28621	.03283
Obs 1 Overrall Rating Domain 3	Formal Observation	35	2.1598	.31847	.05383
	Differentiated Supervision	76	2.1127	.35301	.04049
Obs 1 Rating	Formal Observation	35	2.1447	.31602	.05342
	Differentiated Supervision	76	2.1043	.29401	.03373
Obs 2 Overall Rating Domain 2	Formal Observation	35	2.1340	.19148	.03237
	Differentiated Supervision	76	2.1525	.20311	.02330
Obs 2 Overrall Rating Domain 3	Formal Observation	35	2.2457	.15837	.02677
	Differentiated Supervision	76	2.2196	.22633	.02596
Obs 2 Rating	Formal Observation	35	2.1899	.16164	.02732
	Differentiated Supervision	76	2.1860	.19896	.02282

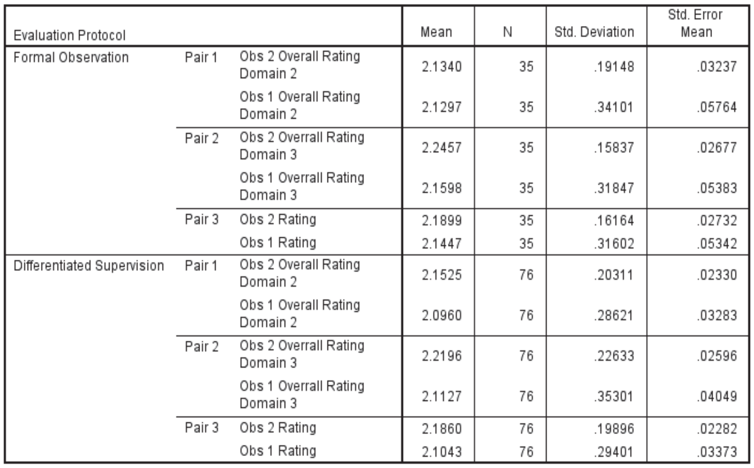
Independent Samples Test



		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Obs 1 Overall Rating Domain 2	Equal variances assumed	1.059	.306	.542	109	.589	.03373	.06217	-.08950	.15695
	Equal variances not assumed			.508	56.921	.613	.03373	.06633	-.09911	.16656
Obs 1 Overrall Rating Domain 3	Equal variances assumed	.933	.336	.673	109	.503	.04708	.06999	-.09163	.18579
	Equal variances not assumed			.699	72.797	.487	.04708	.06736	-.08718	.18134
Obs 1 Rating	Equal variances assumed	.031	.860	.657	109	.513	.04040	.06150	-.08148	.16229
	Equal variances not assumed			.640	62.040	.525	.04040	.06317	-.08588	.16668
Obs 2 Overall Rating Domain 2	Equal variances assumed	.024	.877	-.453	109	.652	-.01846	.04076	-.09925	.06234
	Equal variances not assumed			-.463	69.862	.645	-.01846	.03988	-.09800	.06108
Obs 2 Overrall Rating Domain 3	Equal variances assumed	2.219	.139	.617	109	.539	.02614	.04240	-.05788	.11017
	Equal variances not assumed			.701	91.386	.485	.02614	.03729	-.04793	.10021
Obs 2 Rating	Equal variances assumed	.609	.437	.100	109	.921	.00384	.03843	-.07232	.08000
	Equal variances not assumed			.108	80.279	.914	.00384	.03560	-.06700	.07469

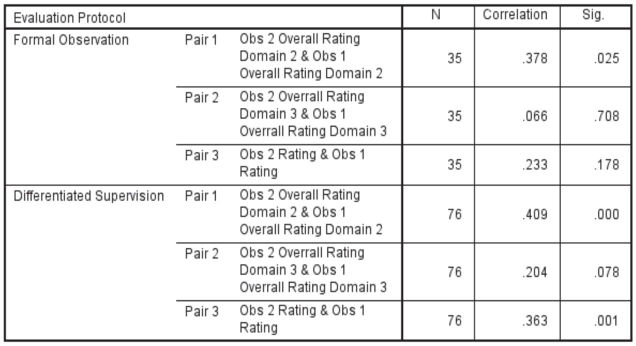
Appendix M

Paired Samples Statistics



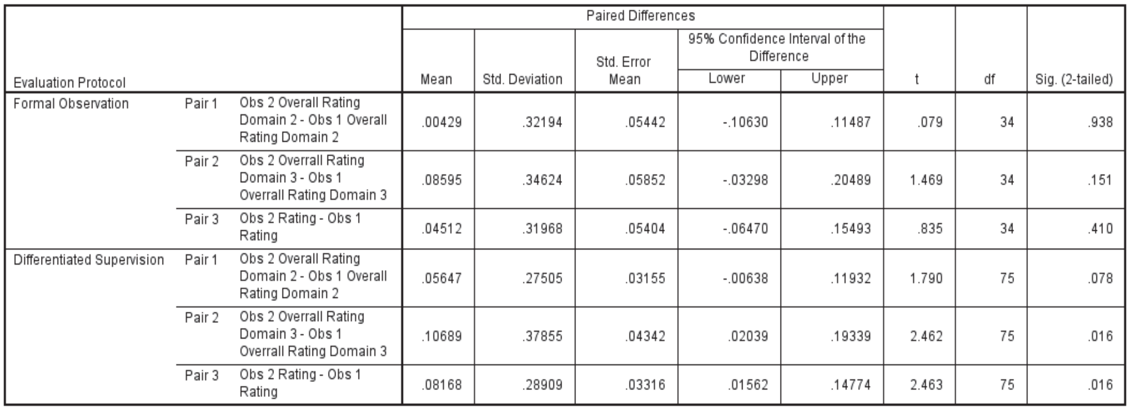
Evaluation Protocol			Mean	N	Std. Deviation	Std. Error Mean
Formal Observation	Pair 1	Obs 2 Overall Rating Domain 2	2.1340	35	.19148	.03237
		Obs 1 Overall Rating Domain 2	2.1297	35	.34101	.05764
	Pair 2	Obs 2 Overrall Rating Domain 3	2.2457	35	.15837	.02677
		Obs 1 Overrall Rating Domain 3	2.1598	35	.31847	.05383
	Pair 3	Obs 2 Rating	2.1899	35	.16164	.02732
		Obs 1 Rating	2.1447	35	.31602	.05342
Differentiated Supervision	Pair 1	Obs 2 Overall Rating Domain 2	2.1525	76	.20311	.02330
		Obs 1 Overall Rating Domain 2	2.0960	76	.28621	.03283
	Pair 2	Obs 2 Overrall Rating Domain 3	2.2196	76	.22633	.02596
		Obs 1 Overrall Rating Domain 3	2.1127	76	.35301	.04049
	Pair 3	Obs 2 Rating	2.1860	76	.19896	.02282
		Obs 1 Rating	2.1043	76	.29401	.03373

Paired Samples Correlations



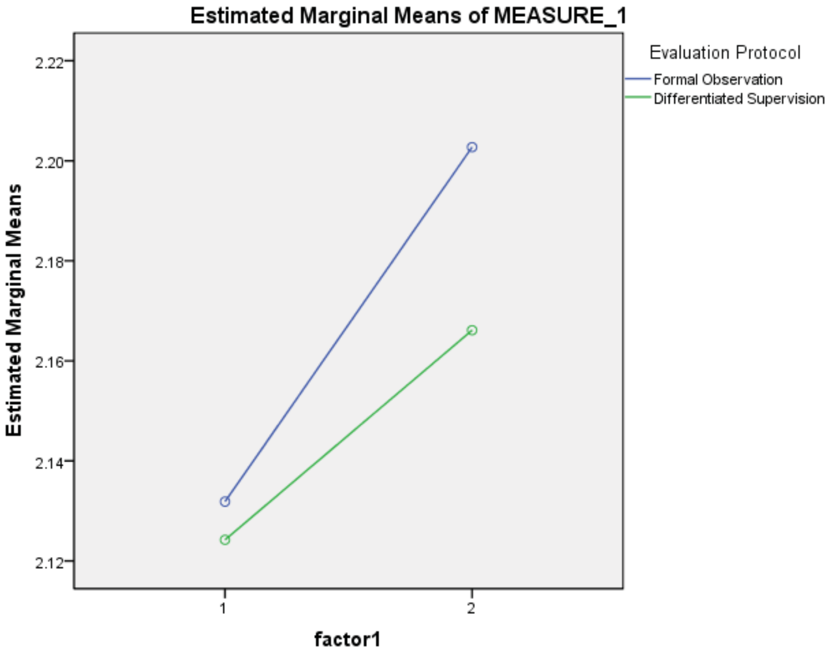
Evaluation Protocol			N	Correlation	Sig.
Formal Observation	Pair 1	Obs 2 Overall Rating Domain 2 & Obs 1 Overall Rating Domain 2	35	.378	.025
	Pair 2	Obs 2 Overrall Rating Domain 3 & Obs 1 Overrall Rating Domain 3	35	.066	.708
	Pair 3	Obs 2 Rating & Obs 1 Rating	35	.233	.178
Differentiated Supervision	Pair 1	Obs 2 Overall Rating Domain 2 & Obs 1 Overall Rating Domain 2	76	.409	.000
	Pair 2	Obs 2 Overrall Rating Domain 3 & Obs 1 Overrall Rating Domain 3	76	.204	.078
	Pair 3	Obs 2 Rating & Obs 1 Rating	76	.363	.001

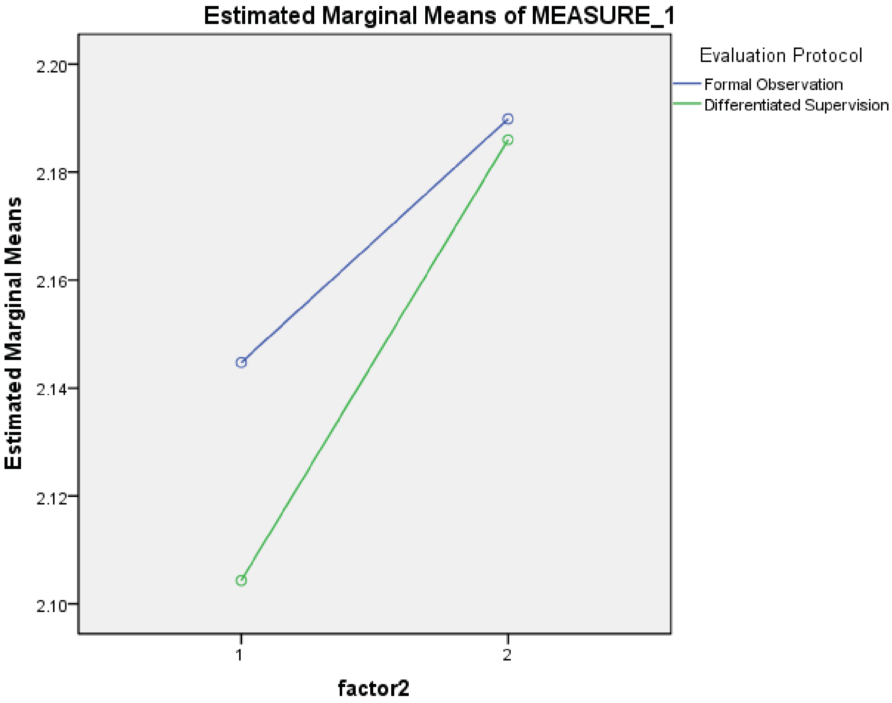
Paired Samples Test



			Paired Differences							
						95% Confidence Interval of the Difference				
Evaluation Protocol			Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Formal Observation	Pair 1	Obs 2 Overall Rating Domain 2 - Obs 1 Overall Rating Domain 2	.00429	.32194	.05442	-.10630	.11487	.079	34	.938
	Pair 2	Obs 2 Overrall Rating Domain 3 - Obs 1 Overrall Rating Domain 3	.08595	.34624	.05852	-.03298	.20489	1.469	34	.151
	Pair 3	Obs 2 Rating - Obs 1 Rating	.04512	.31968	.05404	-.06470	.15493	.835	34	.410
Differentiated Supervision	Pair 1	Obs 2 Overall Rating Domain 2 - Obs 1 Overall Rating Domain 2	.05647	.27505	.03155	-.00638	.11932	1.790	75	.078
	Pair 2	Obs 2 Overrall Rating Domain 3 - Obs 1 Overrall Rating Domain 3	.10689	.37855	.04342	.02039	.19339	2.462	75	.016
	Pair 3	Obs 2 Rating - Obs 1 Rating	.08168	.28909	.03316	.01562	.14774	2.463	75	.016

Appendix N


Estimated Marginal Means of MEASURE_1
Estimated Marginal Means
2.22
2.20
2.18
2.16
2.14
2.12
1
2
factor1
Evaluation Protocol
Formal Observation
Differentiated Supervision


Estimated Marginal Means of MEASURE_1
Estimated Marginal Means
2.20
2.18
2.16
2.14
2.12
2.10
1
2
factor2
Evaluation Protocol
Formal Observation
Differentiated Supervision

Box's Test of Equality of Covariance Matricesa

Box's M	21.046
F	2.002
df1	10
df2	21404.456
Sig.	.029

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

a. Design: Intercept + Protocol

Within Subjects Design: factor1 + factor2 + factor1 * factor2

Tests of Within-Subjects Effects

Measure:MEASURE_1

Source		Sig.	Partial Eta Squared	Noncent. Parameter
factor1	Sphericity Assumed	.000	.122	15.188
	Greenhouse-Geisser	.000	.122	15.188
	Huynh-Feldt	.000	.122	15.188
	Lower-bound	.000	.122	15.188
factor1 * Protocol	Sphericity Assumed	.319	.009	1.002
	Greenhouse-Geisser	.319	.009	1.002
	Huynh-Feldt	.319	.009	1.002
	Lower-bound	.319	.009	1.002
factor2	Sphericity Assumed	.040	.038	4.311
	Greenhouse-Geisser	.040	.038	4.311
	Huynh-Feldt	.040	.038	4.311
	Lower-bound	.040	.038	4.311

factor2 * Protocol	Sphericity Assumed	.551	.003	.358
	Greenhouse-Geisser	.551	.003	.358
	Huynh-Feldt	.551	.003	.358
	Lower-bound	.551	.003	.358
factor1 * factor2	Sphericity Assumed	.027	.044	5.023
	Greenhouse-Geisser	.027	.044	5.023
	Huynh-Feldt	.027	.044	5.023
	Lower-bound	.027	.044	5.023
factor1 * factor2 * Protocol	Sphericity Assumed	.597	.003	.281
	Greenhouse-Geisser	.597	.003	.281
	Huynh-Feldt	.597	.003	.281
	Lower-bound	.597	.003	.281

Tests of Within-Subjects Effects

Measure:MEASURE_1

Source		Observed Power[a]
factor1	Sphericity Assumed	.971
	Greenhouse-Geisser	.971
	Huynh-Feldt	.971
	Lower-bound	.971
factor1 * Protocol	Sphericity Assumed	.168
	Greenhouse-Geisser	.168
	Huynh-Feldt	.168

	Lower-bound	.168
factor2	Sphericity Assumed	.539
	Greenhouse-Geisser	.539
	Huynh-Feldt	.539
	Lower-bound	.539
factor2 * Protocol	Sphericity Assumed	.091
	Greenhouse-Geisser	.091
	Huynh-Feldt	.091
	Lower-bound	.091
factor1 * factor2	Sphericity Assumed	.603
	Greenhouse-Geisser	.603
	Huynh-Feldt	.603
	Lower-bound	.603
factor1 * factor2 * Protocol	Sphericity Assumed	.082
	Greenhouse-Geisser	.082
	Huynh-Feldt	.082
	Lower-bound	.082

a. Computed using alpha = .05

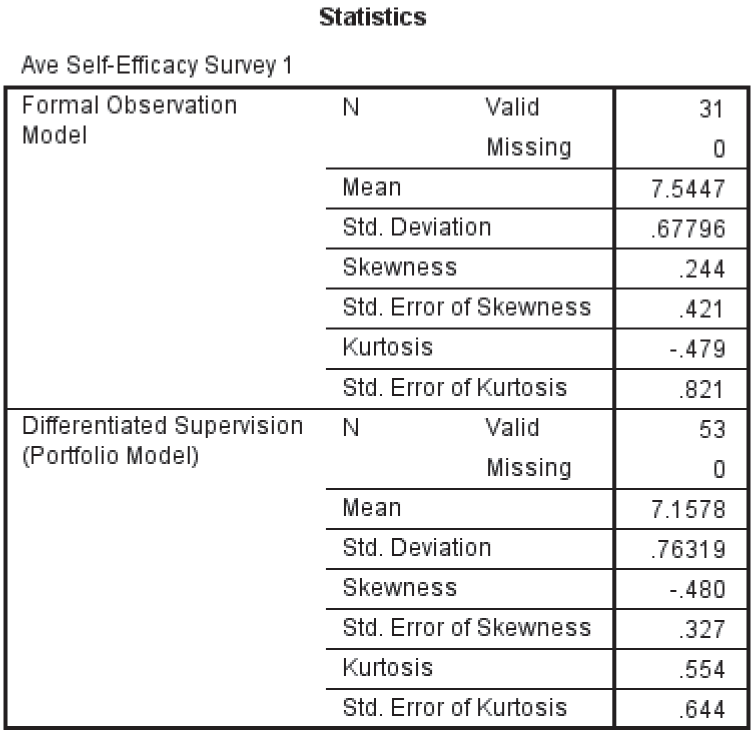
Within-Subjects Effects

Measure:MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F
factor1	Sphericity Assumed	.305	1	.305	15.188
	Greenhouse-Geisser	.305	1.000	.305	15.188
	Huynh-Feldt	.305	1.000	.305	15.188
	Lower-bound	.305	1.000	.305	15.188
factor1 * Protocol	Sphericity Assumed	.020	1	.020	1.002
	Greenhouse-Geisser	.020	1.000	.020	1.002
	Huynh-Feldt	.020	1.000	.020	1.002
	Lower-bound	.020	1.000	.020	1.002
Error(factor1)	Sphericity Assumed	2.188	109	.020	
	Greenhouse-Geisser	2.188	109.000	.020	
	Huynh-Feldt	2.188	109.000	.020	
	Lower-bound	2.188	109.000	.020	

factor2	Sphericity Assumed	.385	1	.385	4.311
	Greenhouse-Geisser	.385	1.000	.385	4.311
	Huynh-Feldt	.385	1.000	.385	4.311
	Lower-bound	.385	1.000	.385	4.311
factor2 * Protocol	Sphericity Assumed	.032	1	.032	.358
	Greenhouse-Geisser	.032	1.000	.032	.358
	Huynh-Feldt	.032	1.000	.032	.358
	Lower-bound	.032	1.000	.032	.358
Error(factor2)	Sphericity Assumed	9.743	109	.089	
	Greenhouse-Geisser	9.743	109.000	.089	
	Huynh-Feldt	9.743	109.000	.089	
	Lower-bound	9.743	109.000	.089	
factor1 * factor2	Sphericity Assumed	.105	1	.105	5.023
	Greenhouse-Geisser	.105	1.000	.105	5.023
	Huynh-Feldt	.105	1.000	.105	5.023
	Lower-bound	.105	1.000	.105	5.023
factor1 * factor2 * Protocol	Sphericity Assumed	.006	1	.006	.281
	Greenhouse-Geisser	.006	1.000	.006	.281
	Huynh-Feldt	.006	1.000	.006	.281
	Lower-bound	.006	1.000	.006	.281
Error(factor1*factor2)	Sphericity Assumed	2.268	109	.021	
	Greenhouse-Geisser	2.268	109.000	.021	
	Huynh-Feldt	2.268	109.000	.021	
	Lower-bound	2.268	109.000	.021	

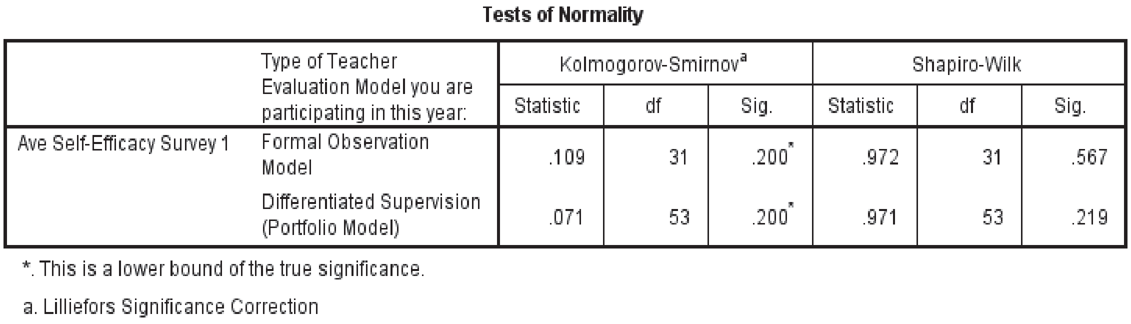
Appendix O



Statistics

Ave Self-Efficacy Survey 1

Formal Observation Model	N	Valid	31
		Missing	0
	Mean		7.5447
	Std. Deviation		.67796
	Skewness		.244
	Std. Error of Skewness		.421
	Kurtosis		-.479
	Std. Error of Kurtosis		.821
Differentiated Supervision (Portfolio Model)	N	Valid	53
		Missing	0
	Mean		7.1578
	Std. Deviation		.76319
	Skewness		-.480
	Std. Error of Skewness		.327
	Kurtosis		.554
	Std. Error of Kurtosis		.644

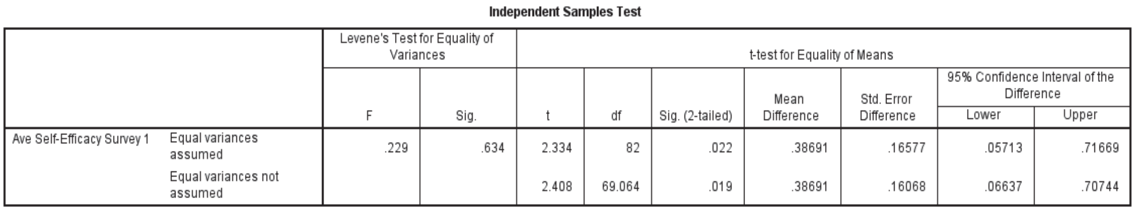


Tests of Normality

	Type of Teacher Evaluation Model you are participating in this year:	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Ave Self-Efficacy Survey 1	Formal Observation Model	.109	31	.200*	.972	31	.567
	Differentiated Supervision (Portfolio Model)	.071	53	.200*	.971	53	.219

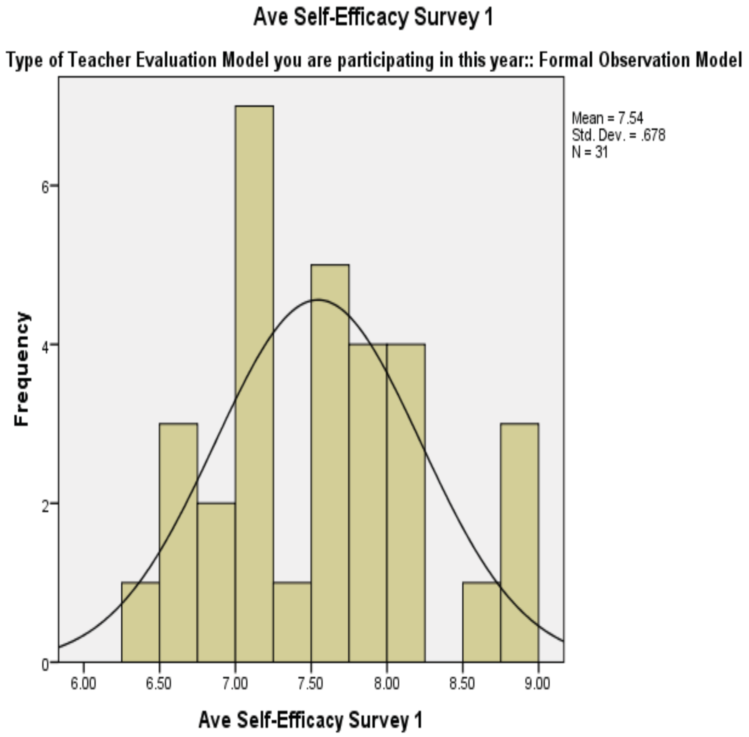
*. This is a lower bound of the true significance.

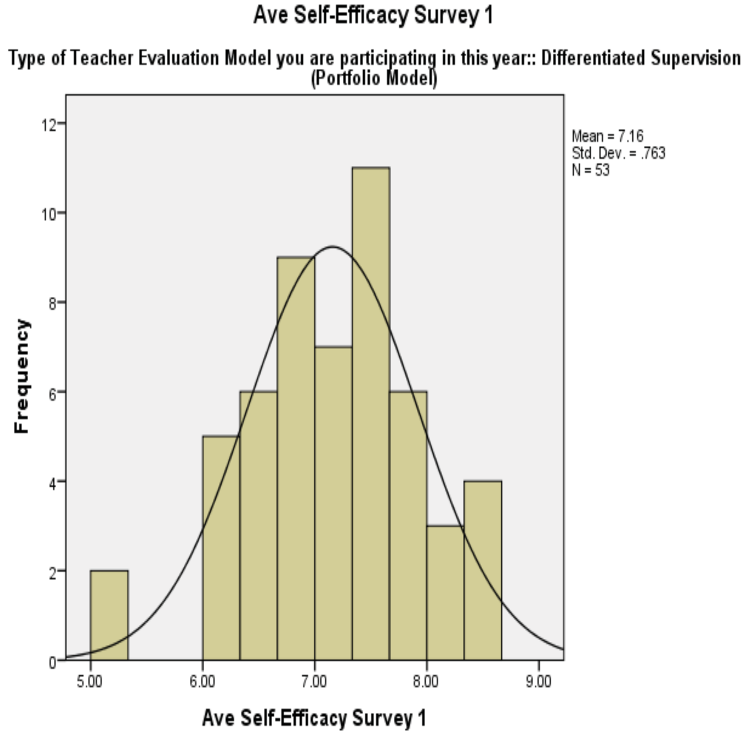
a. Lilliefors Significance Correction



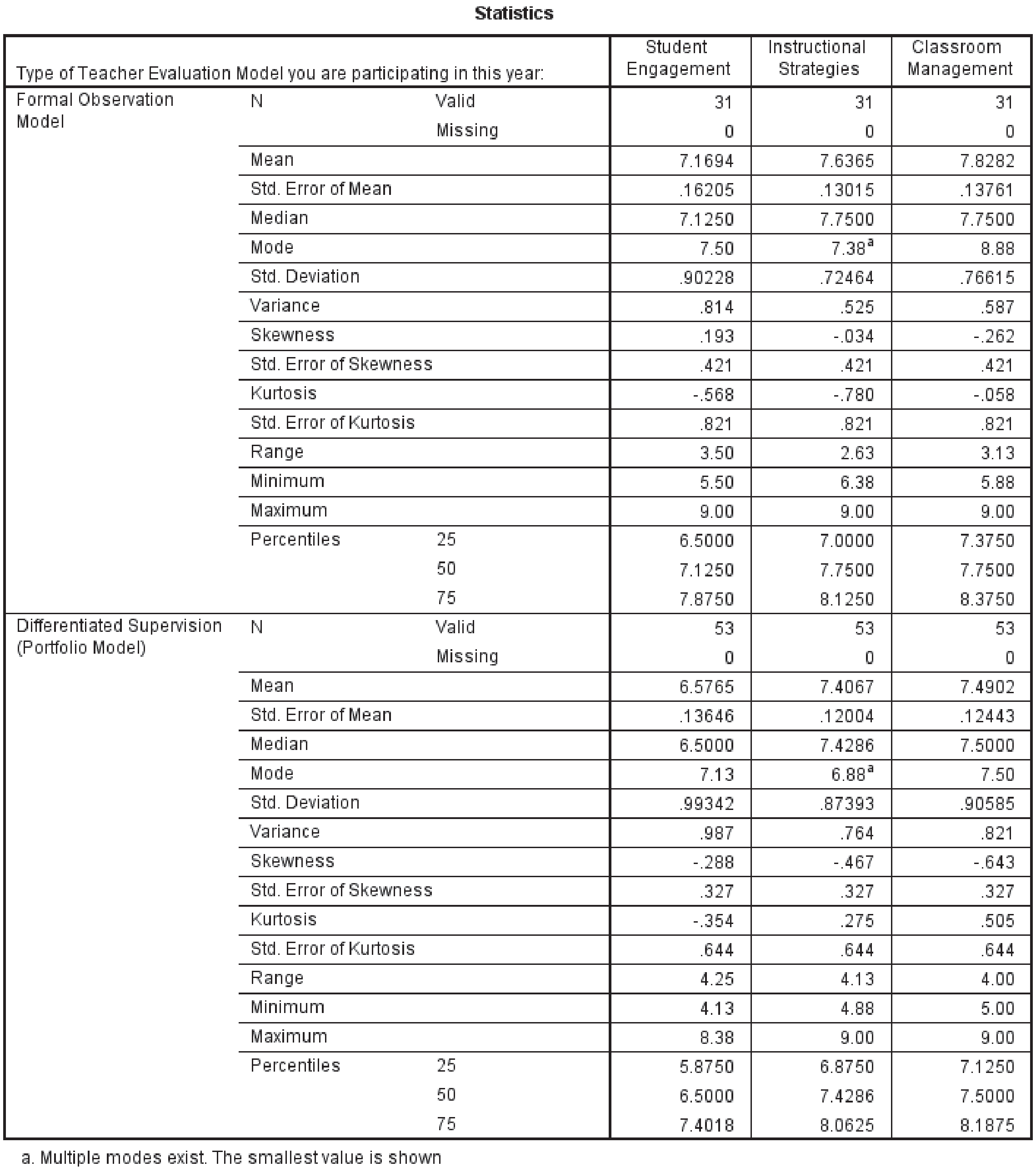
Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Ave Self-Efficacy Survey 1	Equal variances assumed	.229	.634	2.334	82	.022	.38691	.16577	.05713	.71669
	Equal variances not assumed			2.408	69.064	.019	.38691	.16068	.06637	.70744


Ave Self-Efficacy Survey 1
Type of Teacher Evaluation Model you are participating in this year:: Formal Observation Model
Mean = 7.54
Std. Dev. = .678
N = 31
Frequency
0
2
4
6
6.00
6.50
7.00
7.50
8.00
8.50
9.00
Ave Self-Efficacy Survey 1


Ave Self-Efficacy Survey 1
Type of Teacher Evaluation Model you are participating in this year:: Differentiated Supervision (Portfolio Model)
Mean = 7.16
Std. Dev. = .763
N = 53
Frequency
0
2
4
6
8
10
12
5.00
6.00
7.00
8.00
9.00
Ave Self-Efficacy Survey 1

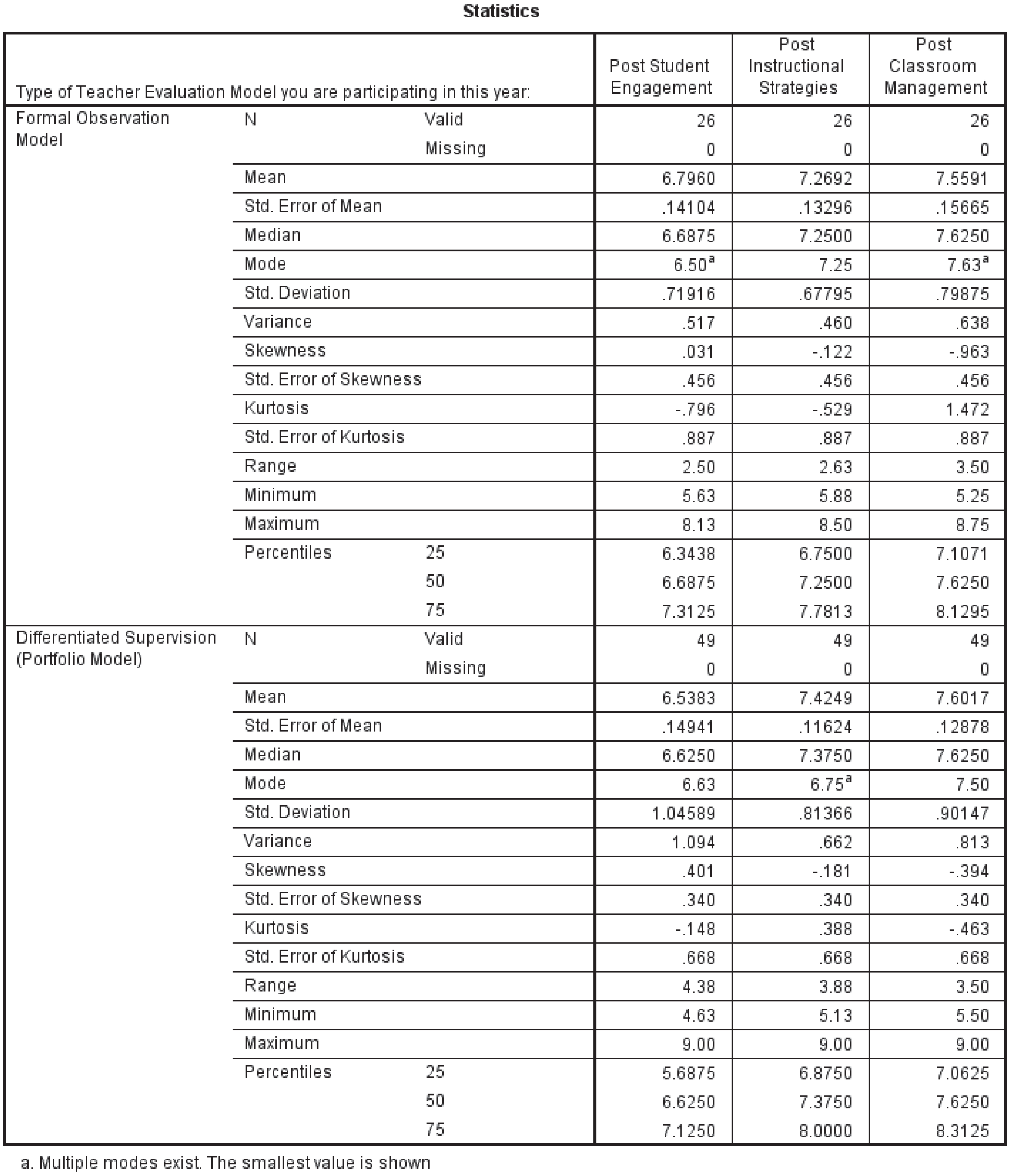
Appendix P



Statistics

Type of Teacher Evaluation Model you are participating in this year:			Student Engagement	Instructional Strategies	Classroom Management
Formal Observation Model	N	Valid	31	31	31
		Missing	0	0	0
	Mean		7.1694	7.6365	7.8282
	Std. Error of Mean		.16205	.13015	.13761
	Median		7.1250	7.7500	7.7500
	Mode		7.50	7.38[a]	8.88
	Std. Deviation		.90228	.72464	.76615
	Variance		.814	.525	.587
	Skewness		.193	-.034	-.262
	Std. Error of Skewness		.421	.421	.421
	Kurtosis		-.568	-.780	-.058
	Std. Error of Kurtosis		.821	.821	.821
	Range		3.50	2.63	3.13
	Minimum		5.50	6.38	5.88
	Maximum		9.00	9.00	9.00
	Percentiles	25	6.5000	7.0000	7.3750
		50	7.1250	7.7500	7.7500
		75	7.8750	8.1250	8.3750
Differentiated Supervision (Portfolio Model)	N	Valid	53	53	53
		Missing	0	0	0
	Mean		6.5765	7.4067	7.4902
	Std. Error of Mean		.13646	.12004	.12443
	Median		6.5000	7.4286	7.5000
	Mode		7.13	6.88[a]	7.50
	Std. Deviation		.99342	.87393	.90585
	Variance		.987	.764	.821
	Skewness		-.288	-.467	-.643
	Std. Error of Skewness		.327	.327	.327
	Kurtosis		-.354	.275	.505
	Std. Error of Kurtosis		.644	.644	.644
	Range		4.25	4.13	4.00
	Minimum		4.13	4.88	5.00
	Maximum		8.38	9.00	9.00
	Percentiles	25	5.8750	6.8750	7.1250
		50	6.5000	7.4286	7.5000
		75	7.4018	8.0625	8.1875

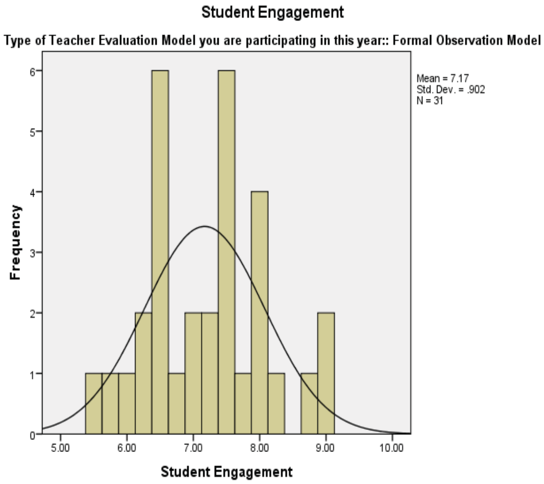
[a] Multiple modes exist. The smallest value is shown

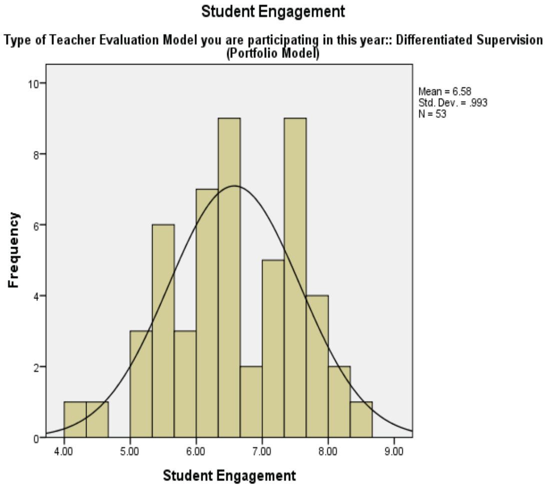


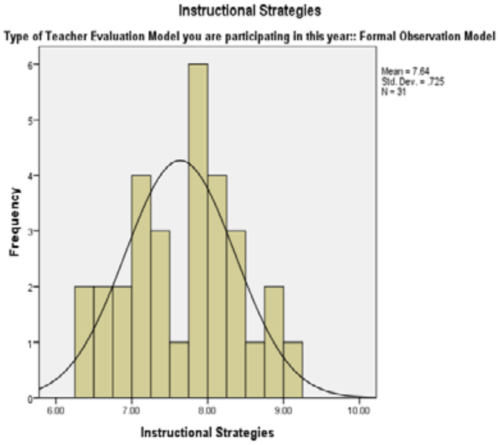
Statistics

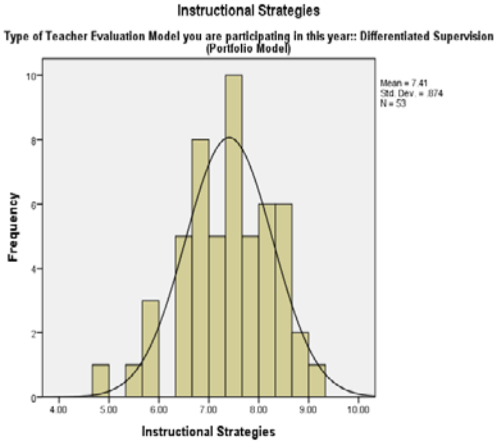
Type of Teacher Evaluation Model you are participating in this year:			Post Student Engagement	Post Instructional Strategies	Post Classroom Management
Formal Observation Model	N	Valid	26	26	26
		Missing	0	0	0
	Mean		6.7960	7.2692	7.5591
	Std. Error of Mean		.14104	.13296	.15665
	Median		6.6875	7.2500	7.6250
	Mode		6.50[a]	7.25	7.63[a]
	Std. Deviation		.71916	.67795	.79875
	Variance		.517	.460	.638
	Skewness		.031	-.122	-.963
	Std. Error of Skewness		.456	.456	.456
	Kurtosis		-.796	-.529	1.472
	Std. Error of Kurtosis		.887	.887	.887
	Range		2.50	2.63	3.50
	Minimum		5.63	5.88	5.25
	Maximum		8.13	8.50	8.75
	Percentiles	25	6.3438	6.7500	7.1071
		50	6.6875	7.2500	7.6250
		75	7.3125	7.7813	8.1295
Differentiated Supervision (Portfolio Model)	N	Valid	49	49	49
		Missing	0	0	0
	Mean		6.5383	7.4249	7.6017
	Std. Error of Mean		.14941	.11624	.12878
	Median		6.6250	7.3750	7.6250
	Mode		6.63	6.75[a]	7.50
	Std. Deviation		1.04589	.81366	.90147
	Variance		1.094	.662	.813
	Skewness		.401	-.181	-.394
	Std. Error of Skewness		.340	.340	.340
	Kurtosis		-.148	.388	-.463
	Std. Error of Kurtosis		.668	.668	.668
	Range		4.38	3.88	3.50
	Minimum		4.63	5.13	5.50
	Maximum		9.00	9.00	9.00
	Percentiles	25	5.6875	6.8750	7.0625
		50	6.6250	7.3750	7.6250
		75	7.1250	8.0000	8.3125

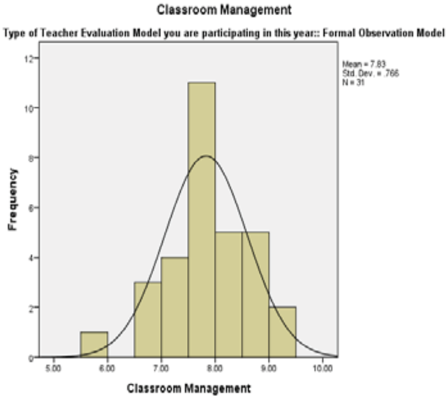
a. Multiple modes exist. The smallest value is shown

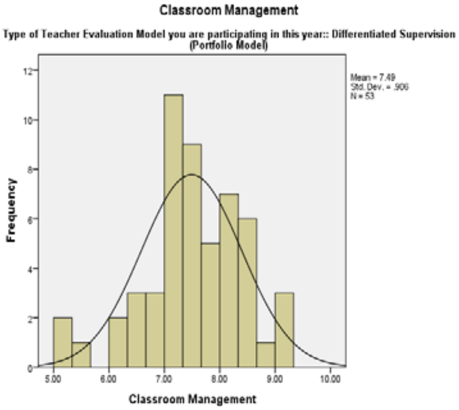

Student Engagement
Type of Teacher Evaluation Model you are participating in this year:: Formal Observation Model
Mean = 7.17
Std. Dev. = .902
N = 31
Frequency
0
1
2
3
4
5
6
5.00
6.00
7.00
8.00
9.00
10.00
Student Engagement

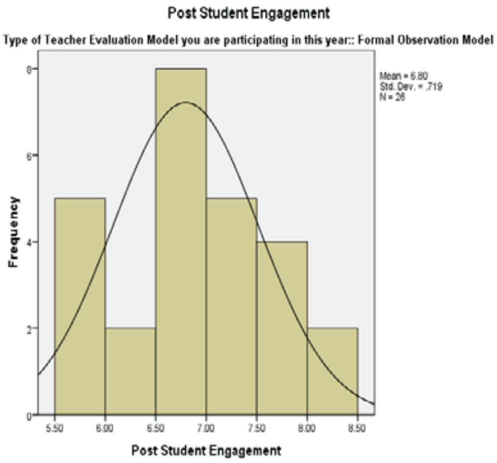

Student Engagement
Type of Teacher Evaluation Model you are participating in this year:: Differentiated Supervision (Portfolio Model)
Mean = 6.58
Std. Dev. = .993
N = 53
Frequency
0
2
4
6
8
10
4.00
5.00
6.00
7.00
8.00
9.00
Student Engagement

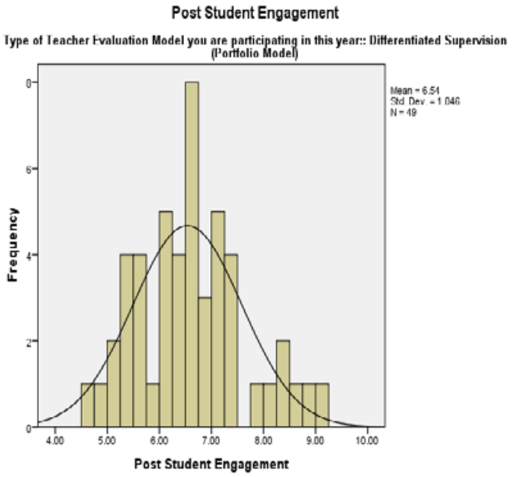

Instructional Strategies
Type of Teacher Evaluation Model you are participating in this year:: Formal Observation Model
Mean = 7.64
Std. Dev. = .725
N = 31
Frequency
0
1
2
3
4
5
6
6.00
7.00
8.00
9.00
10.00
Instructional Strategies

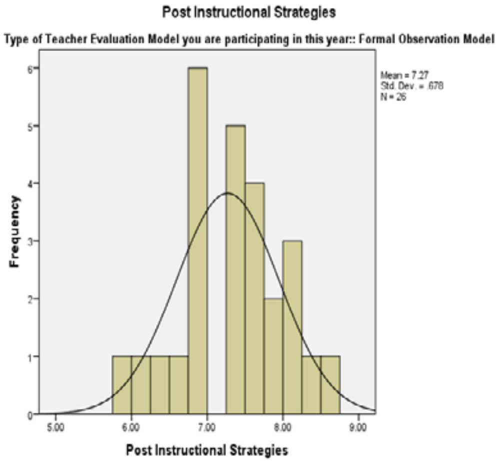

Instructional Strategies
Type of Teacher Evaluation Model you are participating in this year:: Differentiated Supervision (Portfolio Model)
Mean = 7.41
Std. Dev. = .874
N = 53
Frequency
0
2
4
6
8
10
4.00
5.00
6.00
7.00
8.00
9.00
10.00
Instructional Strategies

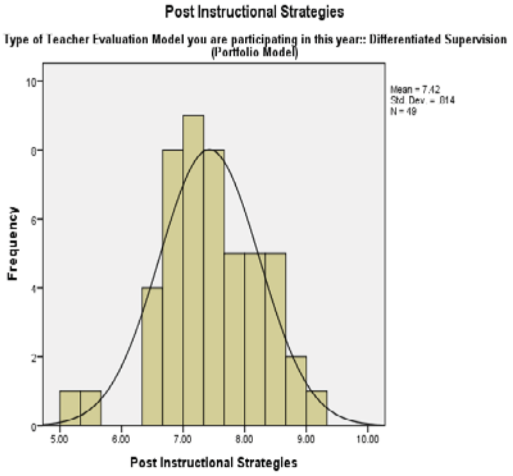

Classroom Management
Type of Teacher Evaluation Model you are participating in this year:: Formal Observation Model
Mean = 7.83
Std. Dev. = .766
N = 31
Frequency
0
2
4
6
8
10
12
5.00
6.00
7.00
8.00
9.00
10.00
Classroom Management

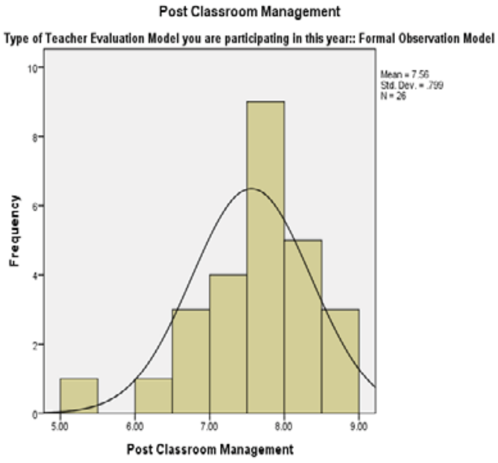

Classroom Management
Type of Teacher Evaluation Model you are participating in this year:: Differentiated Supervision (Portfolio Model)
Mean = 7.49
Std. Dev. = .906
N = 53
Frequency
0
2
4
6
8
10
12
5.00
6.00
7.00
8.00
9.00
10.00
Classroom Management

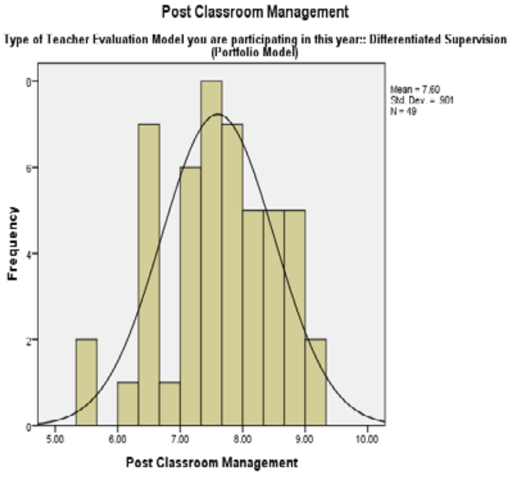

Post Student Engagement
Type of Teacher Evaluation Model you are participating in this year:: Formal Observation Model
Mean = 6.80
Std. Dev. = .719
N = 26
Frequency
0
2
4
6
8
5.50
6.00
6.50
7.00
7.50
8.00
8.50
Post Student Engagement

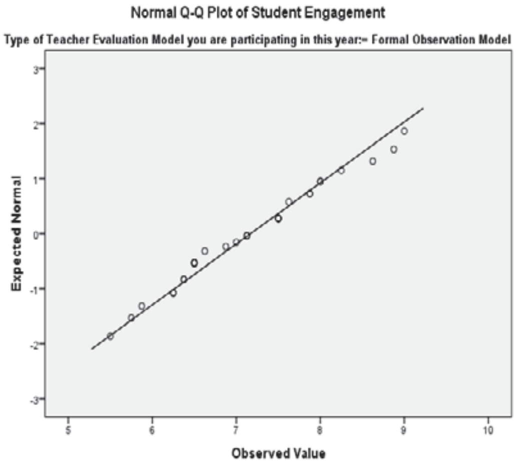

Post Student Engagement
Type of Teacher Evaluation Model you are participating in this year:: Differentiated Supervision (Portfolio Model)
Mean = 6.54
Std. Dev. = 1.046
N = 49
Frequency
0
2
4
6
8
4.00
5.00
6.00
7.00
8.00
9.00
10.00
Post Student Engagement

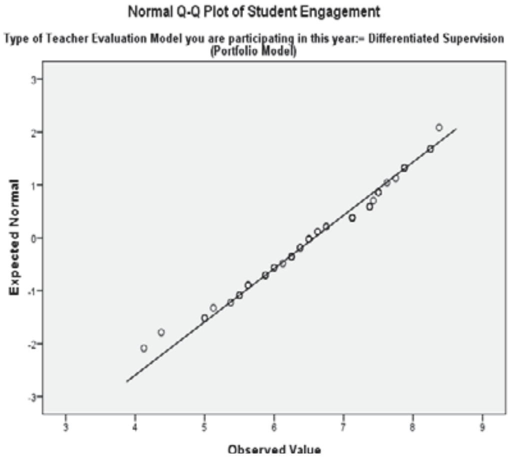

Post Instructional Strategies
Type of Teacher Evaluation Model you are participating in this year:: Formal Observation Model
Mean = 7.27
Std. Dev. = .678
N = 26
Frequency
0
1
2
3
4
5
6
5.00
6.00
7.00
8.00
9.00
Post Instructional Strategies

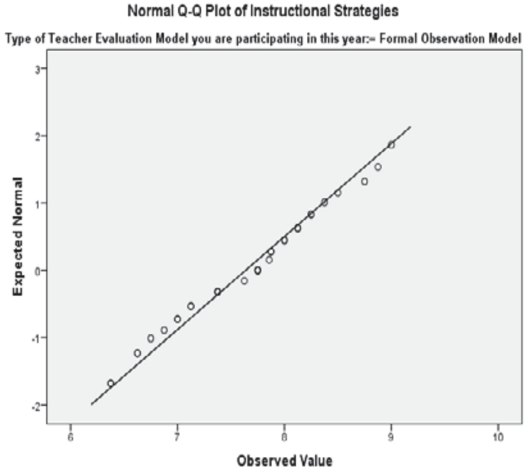

Post Instructional Strategies
Type of Teacher Evaluation Model you are participating in this year:: Differentiated Supervision (Portfolio Model)
Mean = 7.42
Std. Dev. = .814
N = 49
Frequency
0
2
4
6
8
10
5.00
6.00
7.00
8.00
9.00
10.00
Post Instructional Strategies

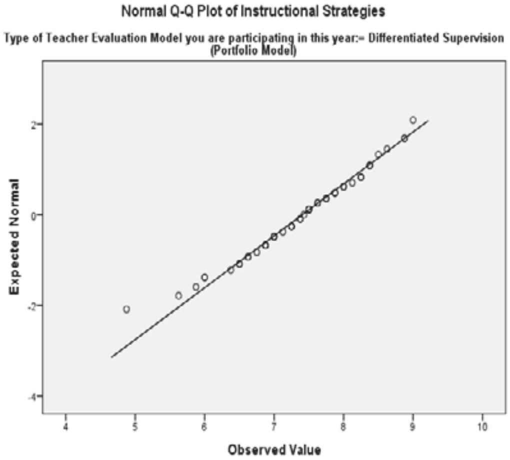

Post Classroom Management
Type of Teacher Evaluation Model you are participating in this year:: Formal Observation Model
Mean = 7.56
Std. Dev. = .799
N = 26
Frequency
0
2
4
6
8
10
5.00
6.00
7.00
8.00
9.00
Post Classroom Management

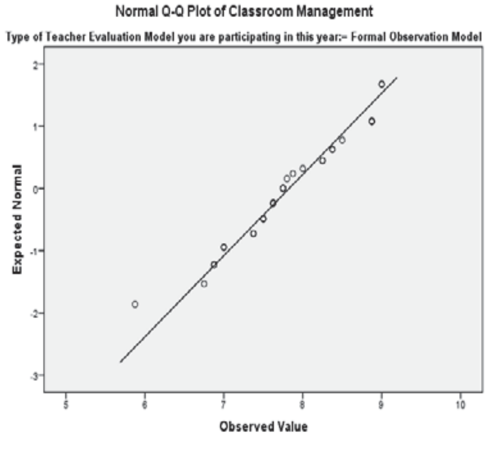

Post Classroom Management
Type of Teacher Evaluation Model you are participating in this year:: Differentiated Supervision (Portfolio Model)
Mean = 7.60
Std. Dev. = .901
N = 49
Frequency
0
2
4
6
8
5.00
6.00
7.00
8.00
9.00
10.00
Post Classroom Management

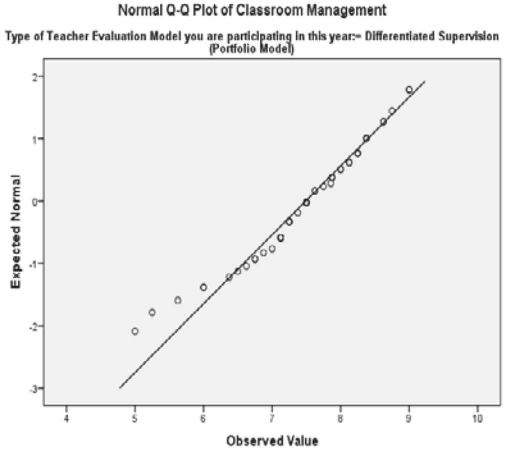

Normal Q-Q Plot of Student Engagement
Type of Teacher Evaluation Model you are participating in this year:= Formal Observation Model
Expected Normal
3
2
1
0
-1
-2
-3
5
6
7
8
9
10
Observed Value

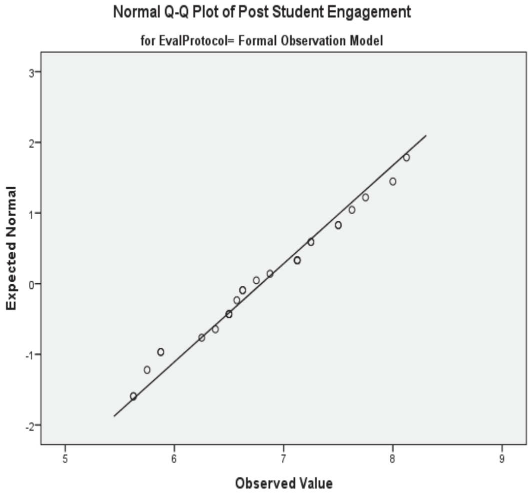

Normal Q-Q Plot of Student Engagement
Type of Teacher Evaluation Model you are participating in this year:= Differentiated Supervision (Portfolio Model)
Expected Normal
3
2
1
0
-1
-2
-3
3
4
5
6
7
8
9
Observed Value

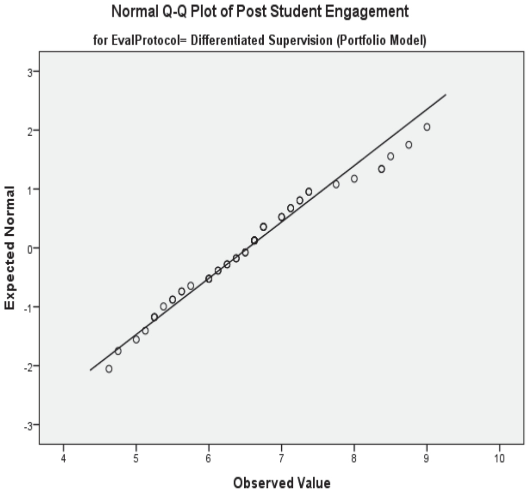

Normal Q-Q Plot of Instructional Strategies
Type of Teacher Evaluation Model you are participating in this year:= Formal Observation Model
Expected Normal
3
2
1
0
-1
-2
6
7
8
9
10
Observed Value

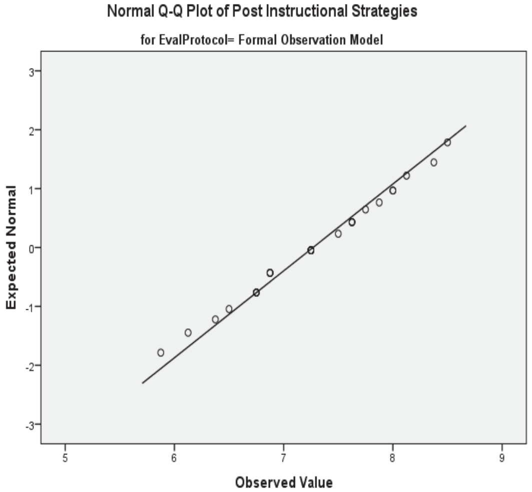

Normal Q-Q Plot of Instructional Strategies
Type of Teacher Evaluation Model you are participating in this year:= Differentiated Supervision (Portfolio Model)
Expected Normal
2
0
-2
-4
4
5
6
7
8
9
10
Observed Value

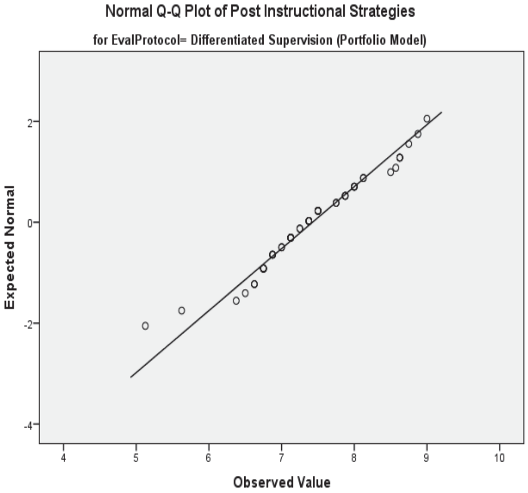

Normal Q-Q Plot of Classroom Management
Type of Teacher Evaluation Model you are participating in this year:= Formal Observation Model
Expected Normal
2
1
0
-1
-2
-3
5
6
7
8
9
10
Observed Value

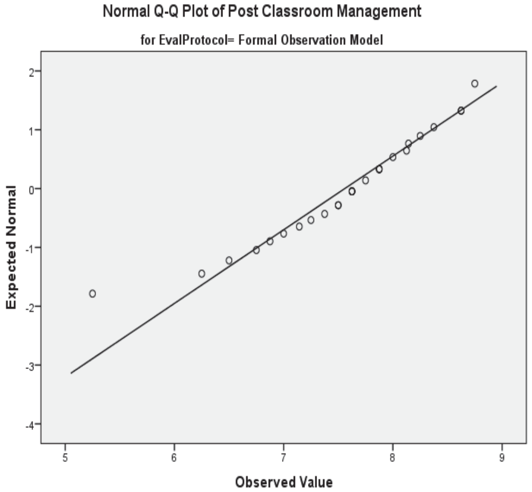

Normal Q-Q Plot of Classroom Management
Type of Teacher Evaluation Model you are participating in this year:= Differentiated Supervision (Portfolio Model)
Expected Normal
2
1
0
-1
-2
-3
4
5
6
7
8
9
10
Observed Value

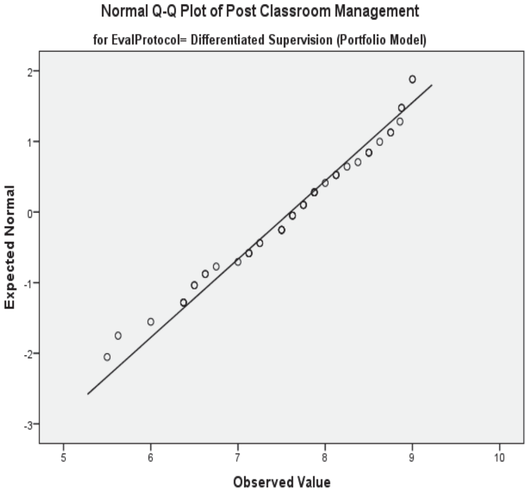

Normal Q-Q Plot of Post Student Engagement
for EvalProtocol= Formal Observation Model
Expected Normal
3
2
1
0
-1
-2
Observed Value
5
6
7
8
9

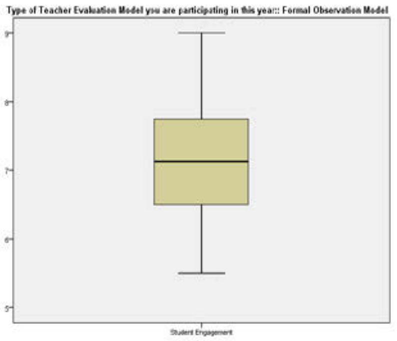

Normal Q-Q Plot of Post Student Engagement
for EvalProtocol= Differentiated Supervision (Portfolio Model)
Expected Normal
3
2
1
0
-1
-2
-3
Observed Value
4
5
6
7
8
9
10

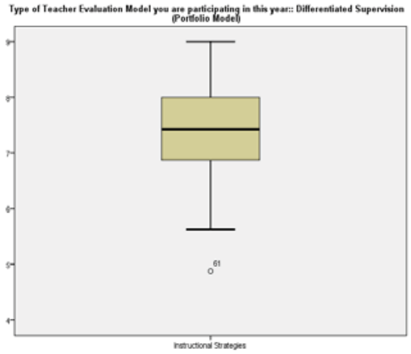

Normal Q-Q Plot of Post Instructional Strategies
for EvalProtocol= Formal Observation Model
Expected Normal
3
2
1
0
-1
-2
-3
Observed Value
5
6
7
8
9

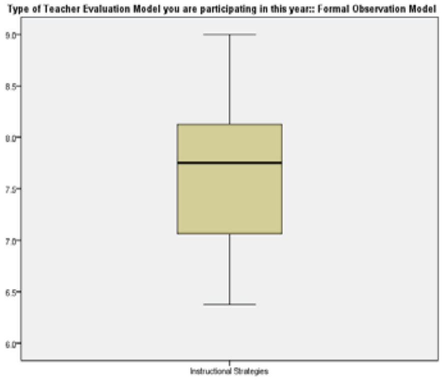

Normal Q-Q Plot of Post Instructional Strategies
for EvalProtocol= Differentiated Supervision (Portfolio Model)
Expected Normal
2
0
-2
-4
Observed Value
4
5
6
7
8
9
10

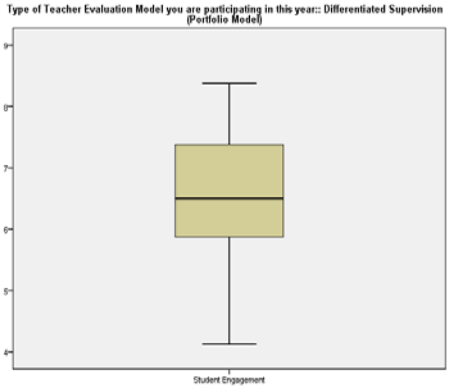

Normal Q-Q Plot of Post Classroom Management
for EvalProtocol= Formal Observation Model
Expected Normal
2
1
0
-1
-2
-3
-4
Observed Value
5
6
7
8
9

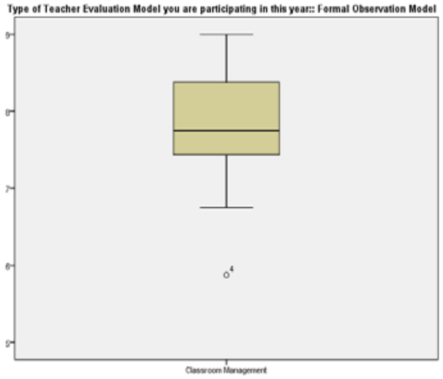

Normal Q-Q Plot of Post Classroom Management
for EvalProtocol= Differentiated Supervision (Portfolio Model)
Expected Normal
2
1
0
-1
-2
-3
Observed Value
5
6
7
8
9
10

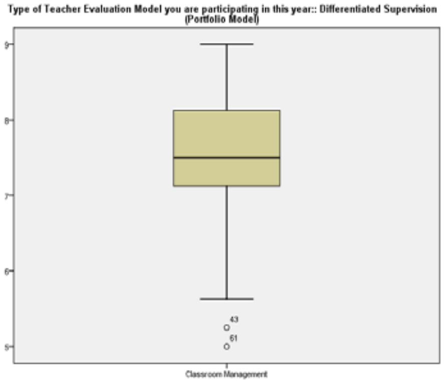

Type of Teacher Evaluation Model you are participating in this year:: Formal Observation Model
9
8
7
6
5
Student Engagement

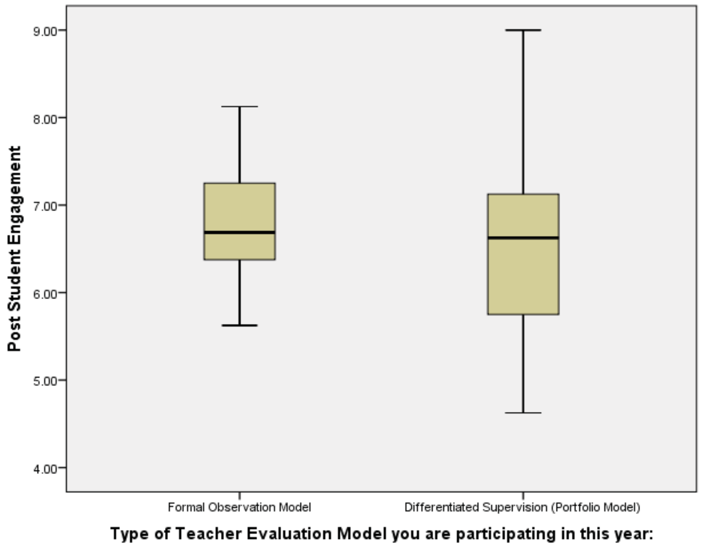

Type of Teacher Evaluation Model you are participating in this year:: Differentiated Supervision (Portfolio Model)
9
8
7
6
5
4
61
Instructional Strategies

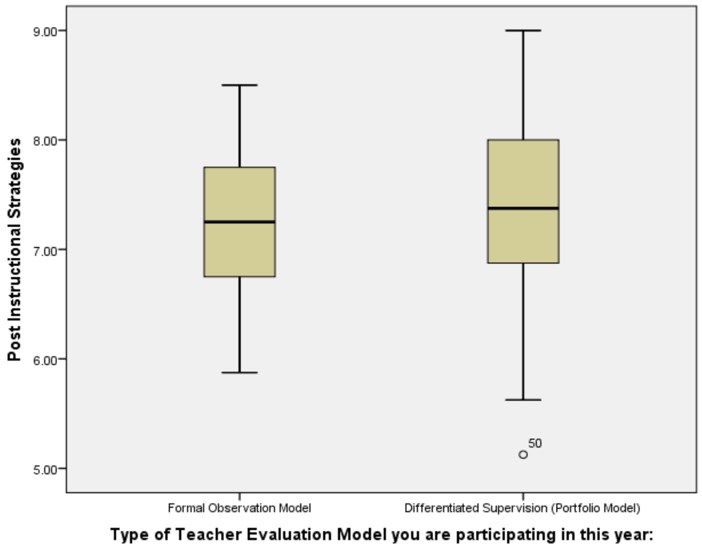

Type of Teacher Evaluation Model you are participating in this year:: Formal Observation Model
9.0
8.5
8.0
7.5
7.0
6.5
6.0
Instructional Strategies

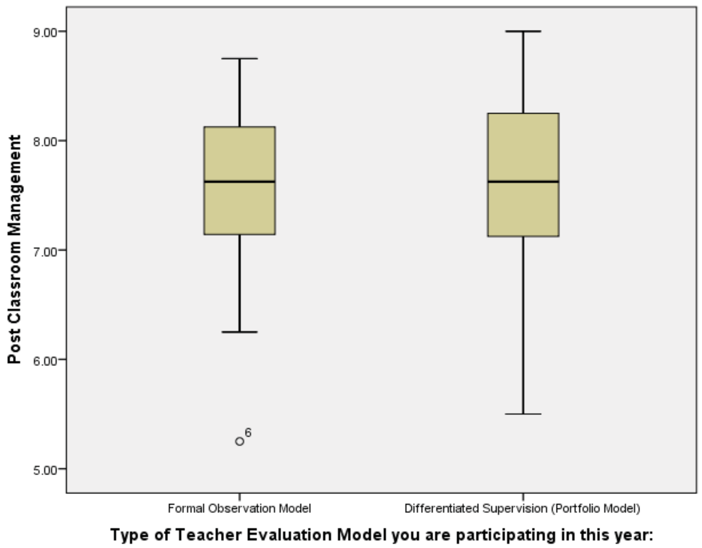

Type of Teacher Evaluation Model you are participating in this year:: Differentiated Supervision (Portfolio Model)
9
8
7
6
5
4
Student Engagement


Type of Teacher Evaluation Model you are participating in this year:: Formal Observation Model
9
8
7
6
5
4
Classroom Management


Type of Teacher Evaluation Model you are participating in this year:: Differentiated Supervision (Portfolio Model)
9
8
7
6
5
43
61
Classroom Management

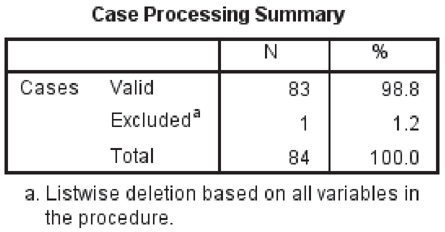

9.00
8.00
7.00
6.00
5.00
4.00
Post Student Engagement
Formal Observation Model
Differentiated Supervision (Portfolio Model)
Type of Teacher Evaluation Model you are participating in this year:


9.00
8.00
7.00
6.00
5.00
Post Instructional Strategies
50
Formal Observation Model
Differentiated Supervision (Portfolio Model)
Type of Teacher Evaluation Model you are participating in this year:


9.00
8.00
7.00
6.00
5.00
Post Classroom Management
6
Formal Observation Model
Differentiated Supervision (Portfolio Model)
Type of Teacher Evaluation Model you are participating in this year:

Appendix Q

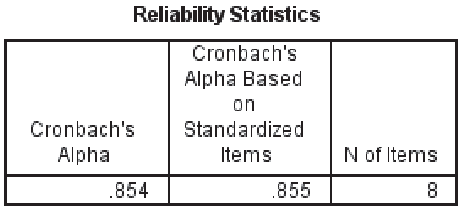
Initial Survey Results – Efficacy in Student Engagement



Case Processing Summary

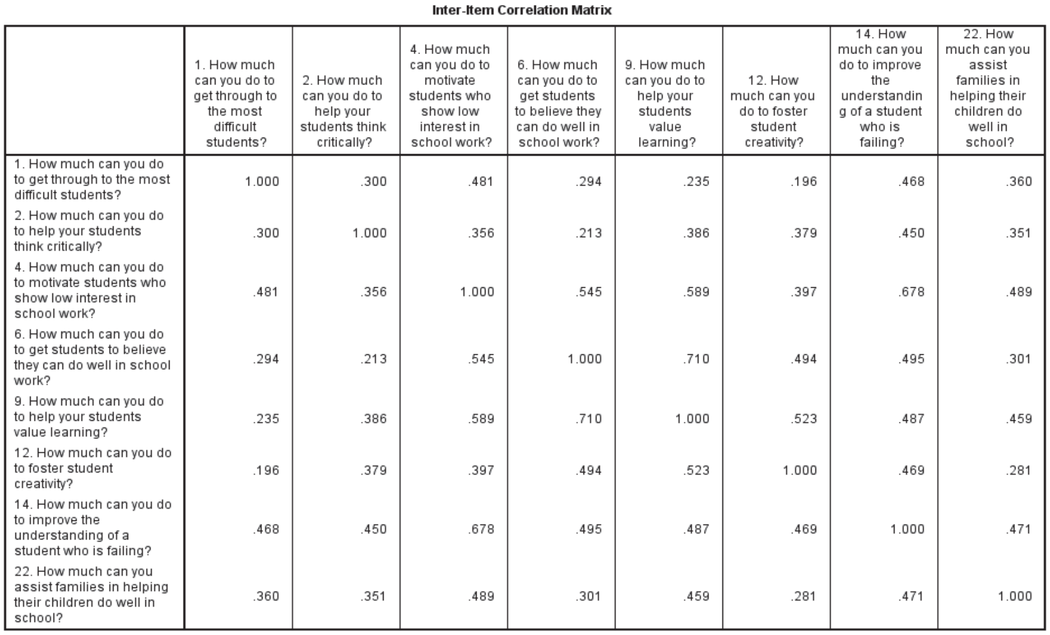
		N	%
Cases	Valid	83	98.8
	Excluded[a]	1	1.2
	Total	84	100.0

a. Listwise deletion based on all variables in the procedure.



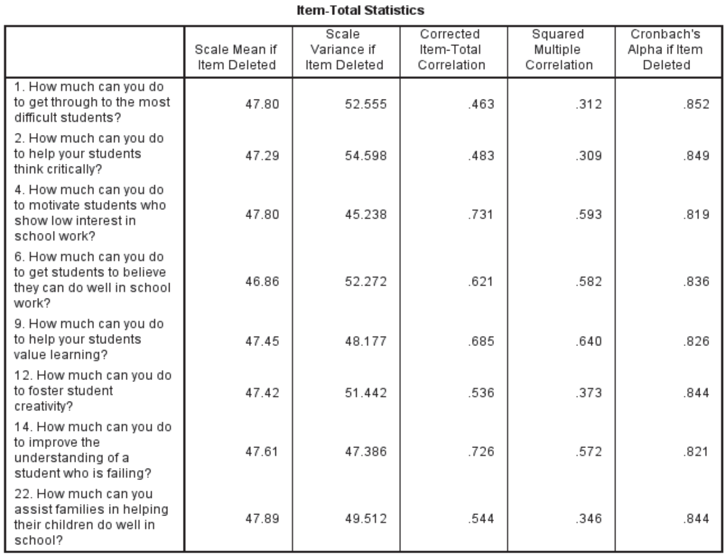
Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.854	.855	8



Inter-Item Correlation Matrix

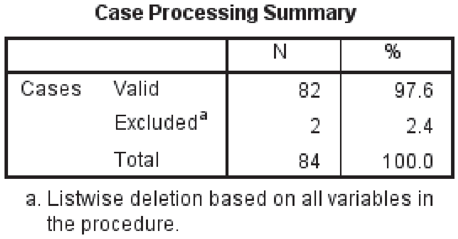
	1. How much can you do to get through to the most difficult students?	2. How much can you do to help your students think critically?	4. How much can you do to motivate students who show low interest in school work?	6. How much can you do to get students to believe they can do well in school work?	9. How much can you do to help your students value learning?	12. How much can you do to foster student creativity?	14. How much can you do to improve the understandin g of a student who is failing?	22. How much can you assist families in helping their children do well in school?
1. How much can you do to get through to the most difficult students?	1.000	.300	.481	.294	.235	.196	.468	.360
2. How much can you do to help your students think critically?	.300	1.000	.356	.213	.386	.379	.450	.351
4. How much can you do to motivate students who show low interest in school work?	.481	.356	1.000	.545	.589	.397	.678	.489
6. How much can you do to get students to believe they can do well in school work?	.294	.213	.545	1.000	.710	.494	.495	.301
9. How much can you do to help your students value learning?	.235	.386	.589	.710	1.000	.523	.487	.459
12. How much can you do to foster student creativity?	.196	.379	.397	.494	.523	1.000	.469	.281
14. How much can you do to improve the understanding of a student who is failing?	.468	.450	.678	.495	.487	.469	1.000	.471
22. How much can you assist families in helping their children do well in school?	.360	.351	.489	.301	.459	.281	.471	1.000



Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
1. How much can you do to get through to the most difficult students?	47.80	52.555	.463	.312	.852
2. How much can you do to help your students think critically?	47.29	54.598	.483	.309	.849
4. How much can you do to motivate students who show low interest in school work?	47.80	45.238	.731	.593	.819
6. How much can you do to get students to believe they can do well in school work?	46.86	52.272	.621	.582	.836
9. How much can you do to help your students value learning?	47.45	48.177	.685	.640	.826
12. How much can you do to foster student creativity?	47.42	51.442	.536	.373	.844
14. How much can you do to improve the understanding of a student who is failing?	47.61	47.386	.726	.572	.821
22. How much can you assist families in helping their children do well in school?	47.89	49.512	.544	.346	.844

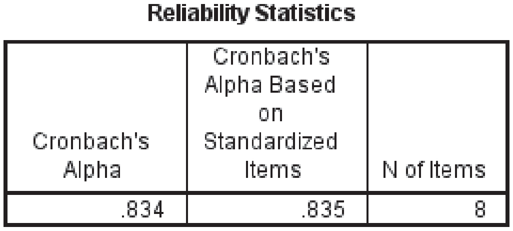
Initial Survey Results – Efficacy in Instructional Strategies



Case Processing Summary

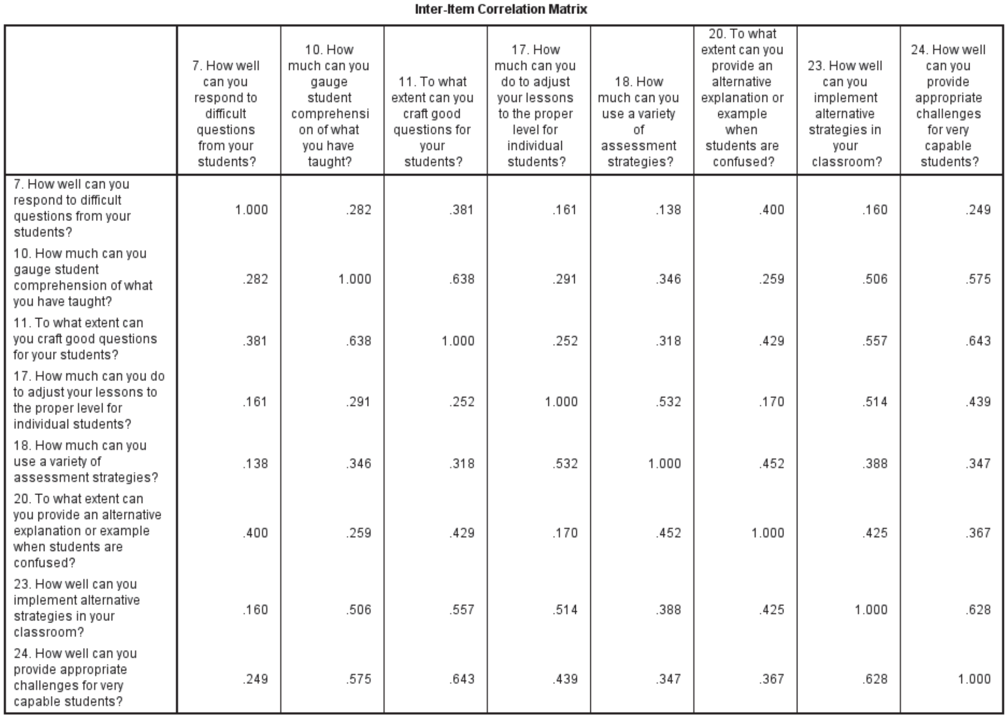
		N	%
Cases	Valid	82	97.6
	Excluded[a]	2	2.4
	Total	84	100.0

a. Listwise deletion based on all variables in the procedure.



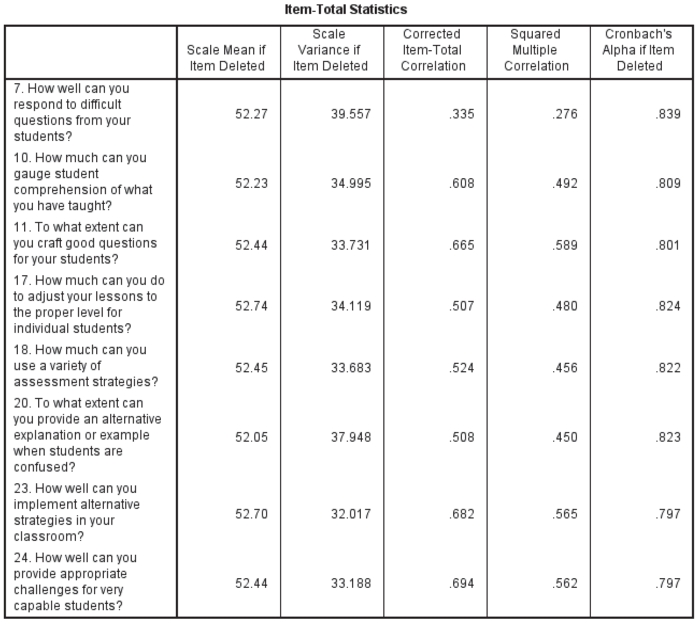
Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.834	.835	8



Inter-Item Correlation Matrix

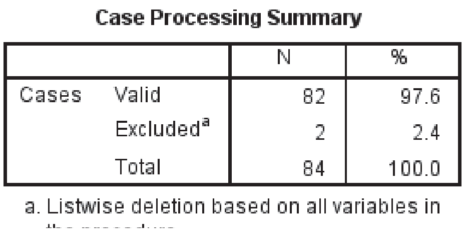
	7. How well can you respond to difficult questions from your students?	10. How much can you gauge student comprehension of what you have taught?	11. To what extent can you craft good questions for your students?	17. How much can you do to adjust your lessons to the proper level for individual students?	18. How much can you use a variety of assessment strategies?	20. To what extent can you provide an alternative explanation or example when students are confused?	23. How well can you implement alternative strategies in your classroom?	24. How well can you provide appropriate challenges for very capable students?
7. How well can you respond to difficult questions from your students?	1.000	.282	.381	.161	.138	.400	.160	.249
10. How much can you gauge student comprehension of what you have taught?	.282	1.000	.638	.291	.346	.259	.506	.575
11. To what extent can you craft good questions for your students?	.381	.638	1.000	.252	.318	.429	.557	.643
17. How much can you do to adjust your lessons to the proper level for individual students?	.161	.291	.252	1.000	.532	.170	.514	.439
18. How much can you use a variety of assessment strategies?	.138	.346	.318	.532	1.000	.452	.388	.347
20. To what extent can you provide an alternative explanation or example when students are confused?	.400	.259	.429	.170	.452	1.000	.425	.367
23. How well can you implement alternative strategies in your classroom?	.160	.506	.557	.514	.388	.425	1.000	.628
24. How well can you provide appropriate challenges for very capable students?	.249	.575	.643	.439	.347	.367	.628	1.000



Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
7. How well can you respond to difficult questions from your students?	52.27	39.557	.335	.276	.839
10. How much can you gauge student comprehension of what you have taught?	52.23	34.995	.608	.492	.809
11. To what extent can you craft good questions for your students?	52.44	33.731	.665	.589	.801
17. How much can you do to adjust your lessons to the proper level for individual students?	52.74	34.119	.507	.480	.824
18. How much can you use a variety of assessment strategies?	52.45	33.683	.524	.456	.822
20. To what extent can you provide an alternative explanation or example when students are confused?	52.05	37.948	.508	.450	.823
23. How well can you implement alternative strategies in your classroom?	52.70	32.017	.682	.565	.797
24. How well can you provide appropriate challenges for very capable students?	52.44	33.188	.694	.562	.797

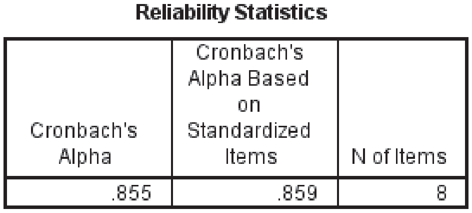
Initial Survey Results – Efficacy in Classroom Management



Case Processing Summary

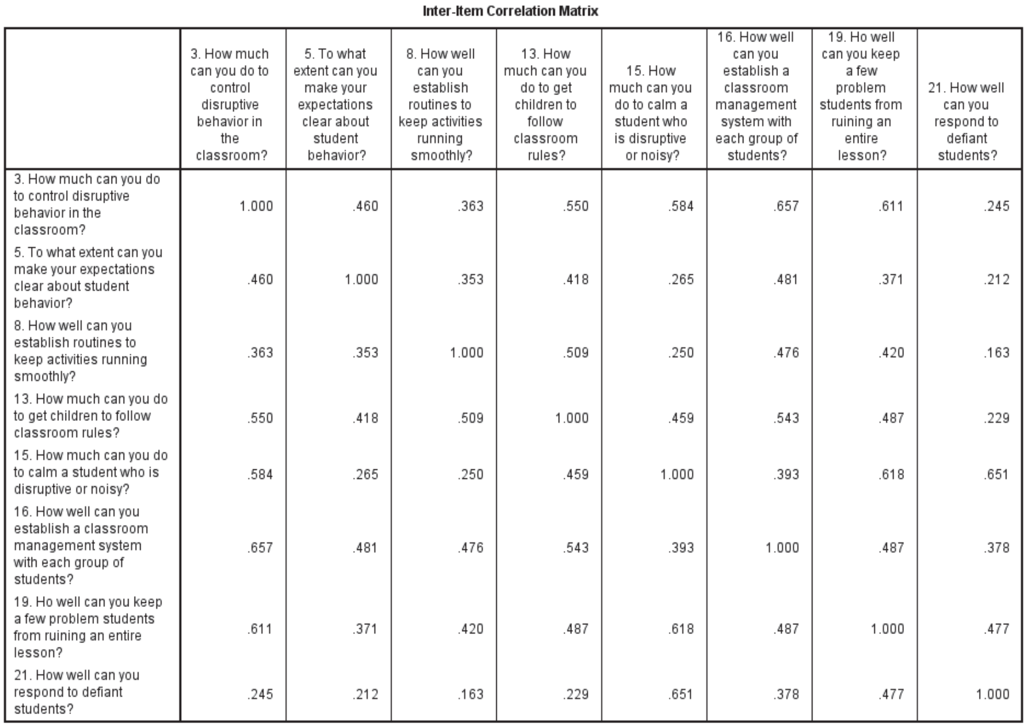
		N	%
Cases	Valid	82	97.6
	Excluded[a]	2	2.4
	Total	84	100.0

a. Listwise deletion based on all variables in



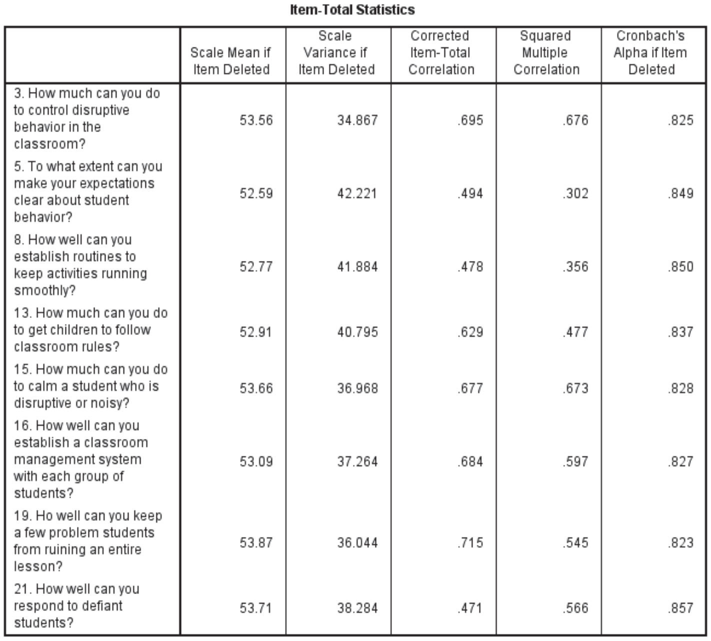
Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.855	.859	8



Inter-Item Correlation Matrix

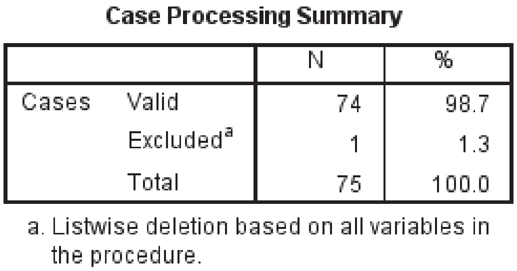
	3. How much can you do to control disruptive behavior in the classroom?	5. To what extent can you make your expectations clear about student behavior?	8. How well can you establish routines to keep activities running smoothly?	13. How much can you do to get children to follow classroom rules?	15. How much can you do to calm a student who is disruptive or noisy?	16. How well can you establish a classroom management system with each group of students?	19. Ho well can you keep a few problem students from ruining an entire lesson?	21. How well can you respond to defiant students?
3. How much can you do to control disruptive behavior in the classroom?	1.000	.460	.363	.550	.584	.657	.611	.245
5. To what extent can you make your expectations clear about student behavior?	.460	1.000	.353	.418	.265	.481	.371	.212
8. How well can you establish routines to keep activities running smoothly?	.363	.353	1.000	.509	.250	.476	.420	.163
13. How much can you do to get children to follow classroom rules?	.550	.418	.509	1.000	.459	.543	.487	.229
15. How much can you do to calm a student who is disruptive or noisy?	.584	.265	.250	.459	1.000	.393	.618	.651
16. How well can you establish a classroom management system with each group of students?	.657	.481	.476	.543	.393	1.000	.487	.378
19. Ho well can you keep a few problem students from ruining an entire lesson?	.611	.371	.420	.487	.618	.487	1.000	.477
21. How well can you respond to defiant students?	.245	.212	.163	.229	.651	.378	.477	1.000



Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
3. How much can you do to control disruptive behavior in the classroom?	53.56	34.867	.695	.676	.825
5. To what extent can you make your expectations clear about student behavior?	52.59	42.221	.494	.302	.849
8. How well can you establish routines to keep activities running smoothly?	52.77	41.884	.478	.356	.850
13. How much can you do to get children to follow classroom rules?	52.91	40.795	.629	.477	.837
15. How much can you do to calm a student who is disruptive or noisy?	53.66	36.968	.677	.673	.828
16. How well can you establish a classroom management system with each group of students?	53.09	37.264	.684	.597	.827
19. Ho well can you keep a few problem students from ruining an entire lesson?	53.87	36.044	.715	.545	.823
21. How well can you respond to defiant students?	53.71	38.284	.471	.566	.857

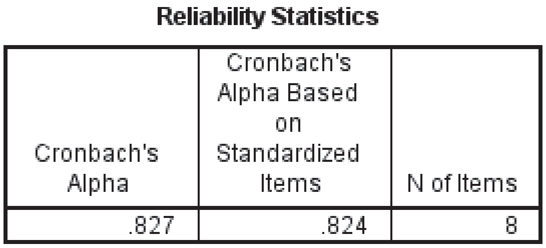
Post Survey Results – Efficacy in Student Engagement



Case Processing Summary

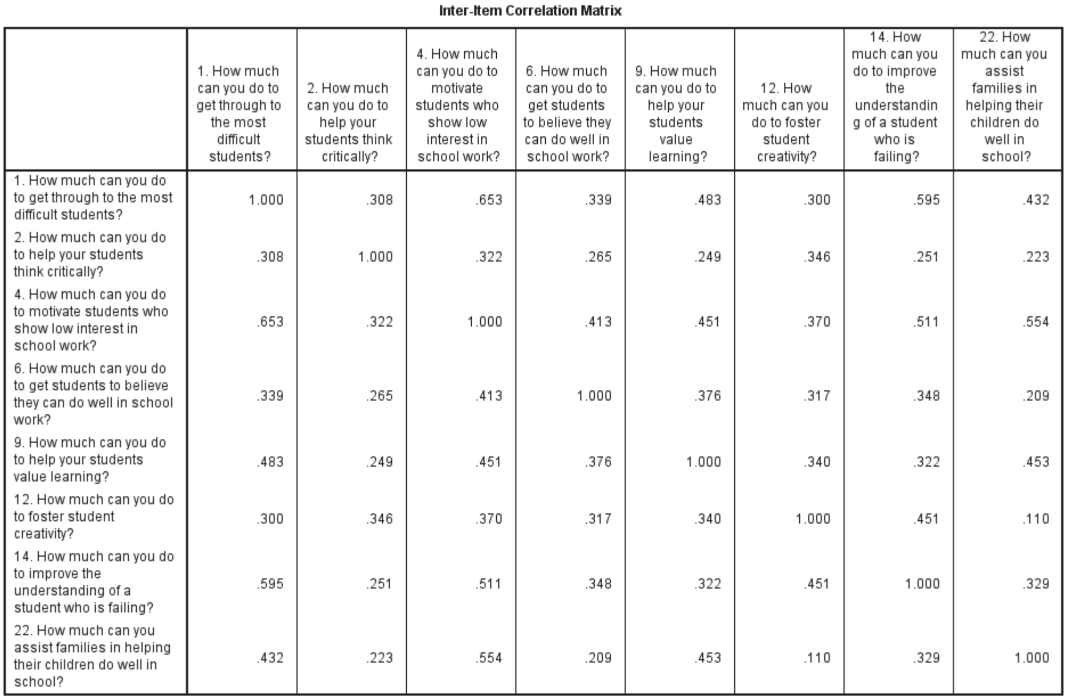
		N	%
Cases	Valid	74	98.7
	Excluded[a]	1	1.3
	Total	75	100.0

a. Listwise deletion based on all variables in the procedure.



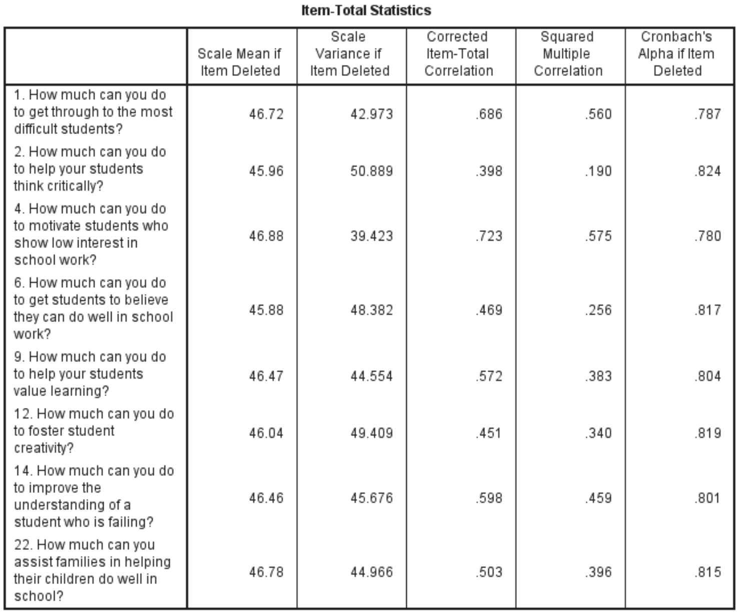
Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.827	.824	8



Inter-Item Correlation Matrix

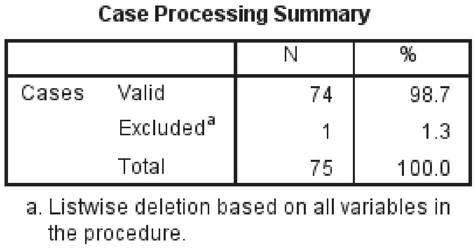
	1. How much can you do to get through to the most difficult students?	2. How much can you do to help your students think critically?	4. How much can you do to motivate students who show low interest in school work?	6. How much can you do to get students to believe they can do well in school work?	9. How much can you do to help your students value learning?	12. How much can you do to foster student creativity?	14. How much can you do to improve the understandin g of a student who is failing?	22. How much can you assist families in helping their children do well in school?
1. How much can you do to get through to the most difficult students?	1.000	.308	.653	.339	.483	.300	.595	.432
2. How much can you do to help your students think critically?	.308	1.000	.322	.265	.249	.346	.251	.223
4. How much can you do to motivate students who show low interest in school work?	.653	.322	1.000	.413	.451	.370	.511	.554
6. How much can you do to get students to believe they can do well in school work?	.339	.265	.413	1.000	.376	.317	.348	.209
9. How much can you do to help your students value learning?	.483	.249	.451	.376	1.000	.340	.322	.453
12. How much can you do to foster student creativity?	.300	.346	.370	.317	.340	1.000	.451	.110
14. How much can you do to improve the understanding of a student who is failing?	.595	.251	.511	.348	.322	.451	1.000	.329
22. How much can you assist families in helping their children do well in school?	.432	.223	.554	.209	.453	.110	.329	1.000



Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
1. How much can you do to get through to the most difficult students?	46.72	42.973	.686	.560	.787
2. How much can you do to help your students think critically?	45.96	50.889	.398	.190	.824
4. How much can you do to motivate students who show low interest in school work?	46.88	39.423	.723	.575	.780
6. How much can you do to get students to believe they can do well in school work?	45.88	48.382	.469	.256	.817
9. How much can you do to help your students value learning?	46.47	44.554	.572	.383	.804
12. How much can you do to foster student creativity?	46.04	49.409	.451	.340	.819
14. How much can you do to improve the understanding of a student who is failing?	46.46	45.676	.598	.459	.801
22. How much can you assist families in helping their children do well in school?	46.78	44.966	.503	.396	.815

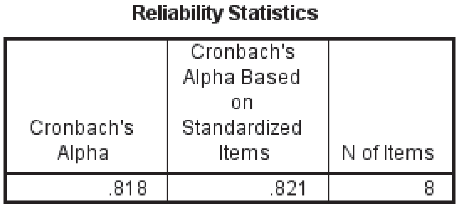
Post Survey Results – Efficacy in Instructional Strategies



Case Processing Summary

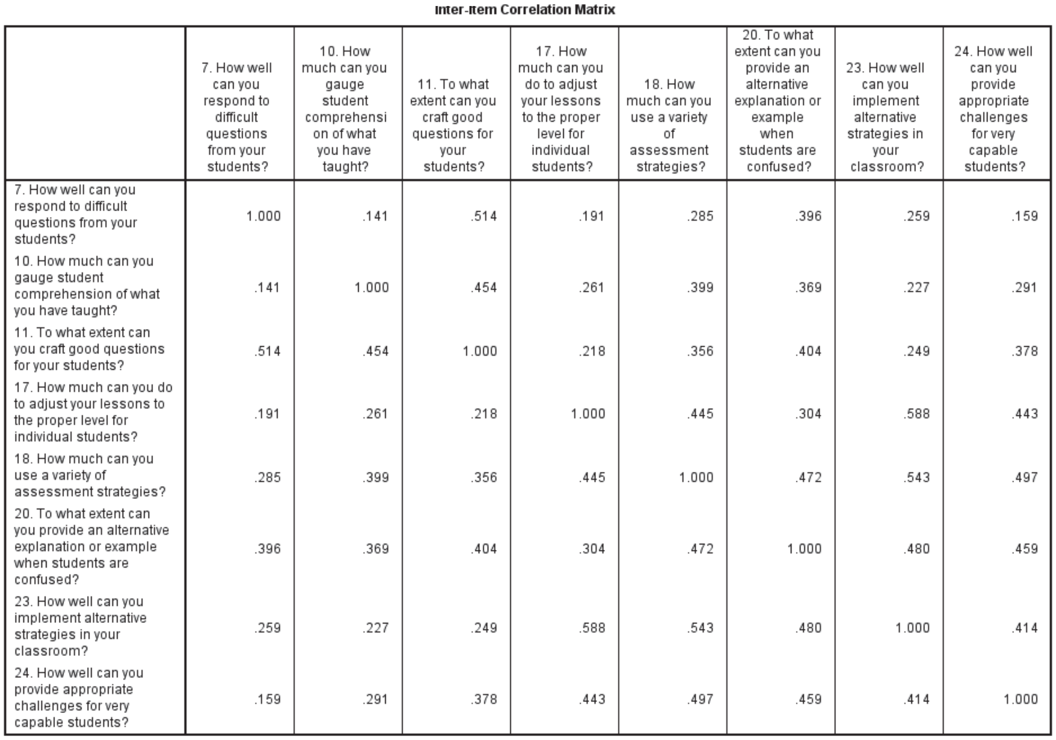
		N	%
Cases	Valid	74	98.7
	Excluded[a]	1	1.3
	Total	75	100.0

a. Listwise deletion based on all variables in the procedure.



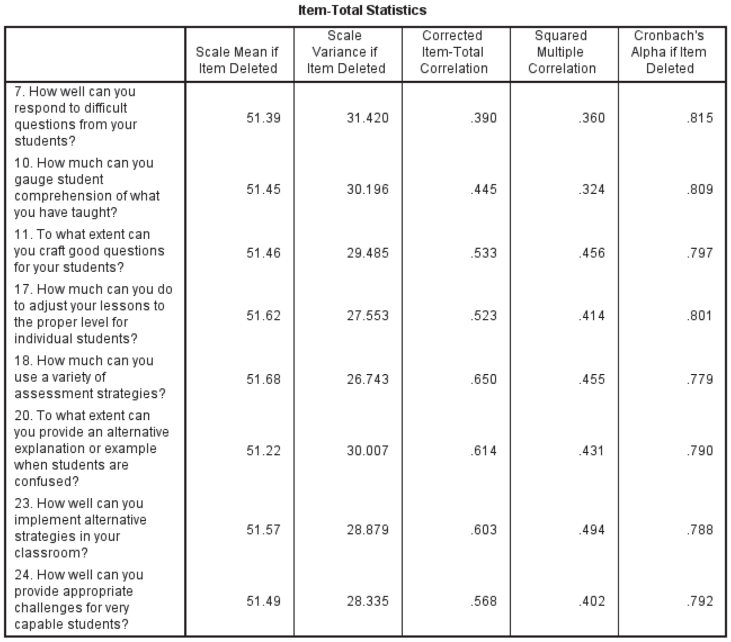
Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.818	.821	8



Inter-Item Correlation Matrix

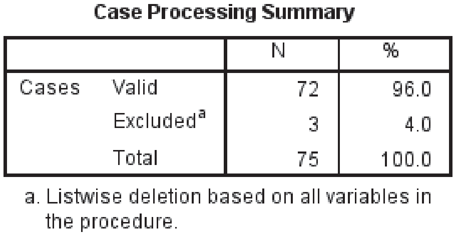
	7. How well can you respond to difficult questions from your students?	10. How much can you gauge student comprehension of what you have taught?	11. To what extent can you craft good questions for your students?	17. How much can you do to adjust your lessons to the proper level for individual students?	18. How much can you use a variety of assessment strategies?	20. To what extent can you provide an alternative explanation or example when students are confused?	23. How well can you implement alternative strategies in your classroom?	24. How well can you provide appropriate challenges for very capable students?
7. How well can you respond to difficult questions from your students?	1.000	.141	.514	.191	.285	.396	.259	.159
10. How much can you gauge student comprehension of what you have taught?	.141	1.000	.454	.261	.399	.369	.227	.291
11. To what extent can you craft good questions for your students?	.514	.454	1.000	.218	.356	.404	.249	.378
17. How much can you do to adjust your lessons to the proper level for individual students?	.191	.261	.218	1.000	.445	.304	.588	.443
18. How much can you use a variety of assessment strategies?	.285	.399	.356	.445	1.000	.472	.543	.497
20. To what extent can you provide an alternative explanation or example when students are confused?	.396	.369	.404	.304	.472	1.000	.480	.459
23. How well can you implement alternative strategies in your classroom?	.259	.227	.249	.588	.543	.480	1.000	.414
24. How well can you provide appropriate challenges for very capable students?	.159	.291	.378	.443	.497	.459	.414	1.000



Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
7. How well can you respond to difficult questions from your students?	51.39	31.420	.390	.360	.815
10. How much can you gauge student comprehension of what you have taught?	51.45	30.196	.445	.324	.809
11. To what extent can you craft good questions for your students?	51.46	29.485	.533	.456	.797
17. How much can you do to adjust your lessons to the proper level for individual students?	51.62	27.553	.523	.414	.801
18. How much can you use a variety of assessment strategies?	51.68	26.743	.650	.455	.779
20. To what extent can you provide an alternative explanation or example when students are confused?	51.22	30.007	.614	.431	.790
23. How well can you implement alternative strategies in your classroom?	51.57	28.879	.603	.494	.788
24. How well can you provide appropriate challenges for very capable students?	51.49	28.335	.568	.402	.792

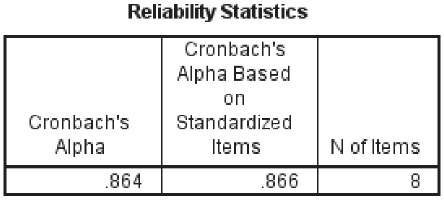
Post Survey Results – Efficacy in Classroom Management



Case Processing Summary

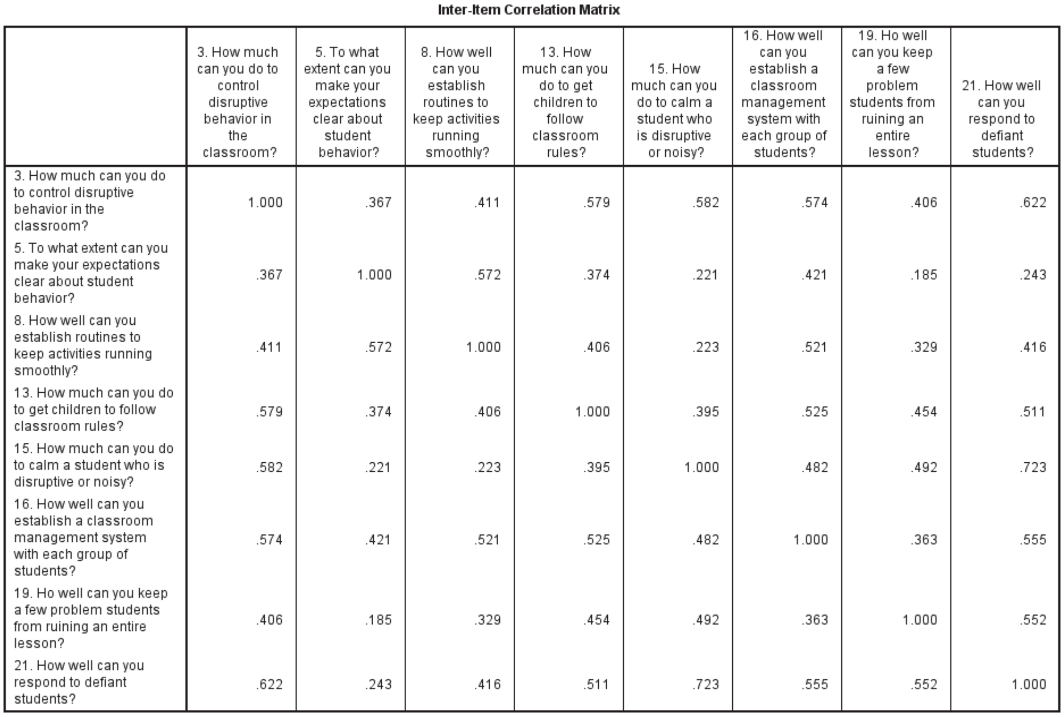
		N	%
Cases	Valid	72	96.0
	Excluded[a]	3	4.0
	Total	75	100.0

a. Listwise deletion based on all variables in the procedure.



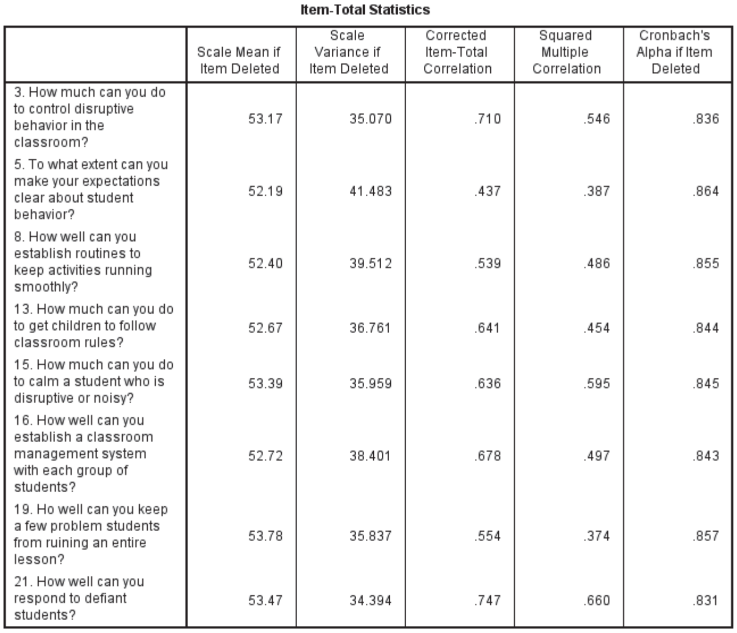
Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.864	.866	8



Inter-Item Correlation Matrix

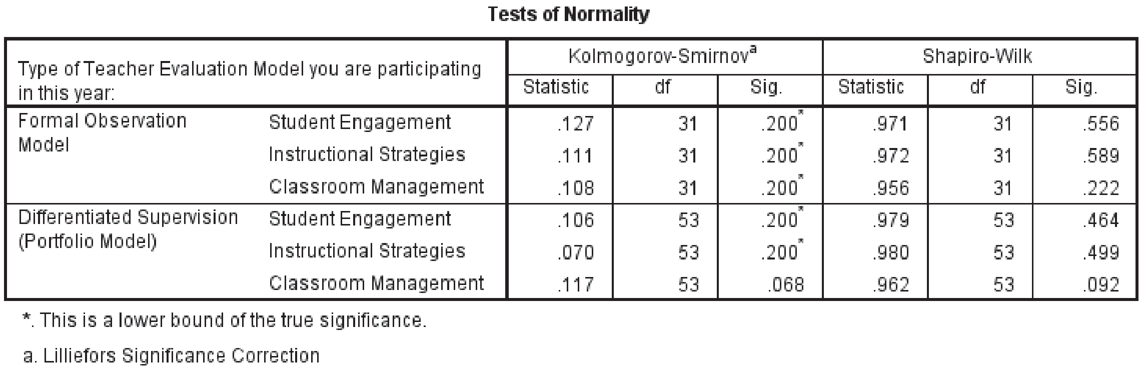
	3. How much can you do to control disruptive behavior in the classroom?	5. To what extent can you make your expectations clear about student behavior?	8. How well can you establish routines to keep activities running smoothly?	13. How much can you do to get children to follow classroom rules?	15. How much can you do to calm a student who is disruptive or noisy?	16. How well can you establish a classroom management system with each group of students?	19. Ho well can you keep a few problem students from ruining an entire lesson?	21. How well can you respond to defiant students?
3. How much can you do to control disruptive behavior in the classroom?	1.000	.367	.411	.579	.582	.574	.406	.622
5. To what extent can you make your expectations clear about student behavior?	.367	1.000	.572	.374	.221	.421	.185	.243
8. How well can you establish routines to keep activities running smoothly?	.411	.572	1.000	.406	.223	.521	.329	.416
13. How much can you do to get children to follow classroom rules?	.579	.374	.406	1.000	.395	.525	.454	.511
15. How much can you do to calm a student who is disruptive or noisy?	.582	.221	.223	.395	1.000	.482	.492	.723
16. How well can you establish a classroom management system with each group of students?	.574	.421	.521	.525	.482	1.000	.363	.555
19. Ho well can you keep a few problem students from ruining an entire lesson?	.406	.185	.329	.454	.492	.363	1.000	.552
21. How well can you respond to defiant students?	.622	.243	.416	.511	.723	.555	.552	1.000



Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
3. How much can you do to control disruptive behavior in the classroom?	53.17	35.070	.710	.546	.836
5. To what extent can you make your expectations clear about student behavior?	52.19	41.483	.437	.387	.864
8. How well can you establish routines to keep activities running smoothly?	52.40	39.512	.539	.486	.855
13. How much can you do to get children to follow classroom rules?	52.67	36.761	.641	.454	.844
15. How much can you do to calm a student who is disruptive or noisy?	53.39	35.959	.636	.595	.845
16. How well can you establish a classroom management system with each group of students?	52.72	38.401	.678	.497	.843
19. Ho well can you keep a few problem students from ruining an entire lesson?	53.78	35.837	.554	.374	.857
21. How well can you respond to defiant students?	53.47	34.394	.747	.660	.831

Appendix R

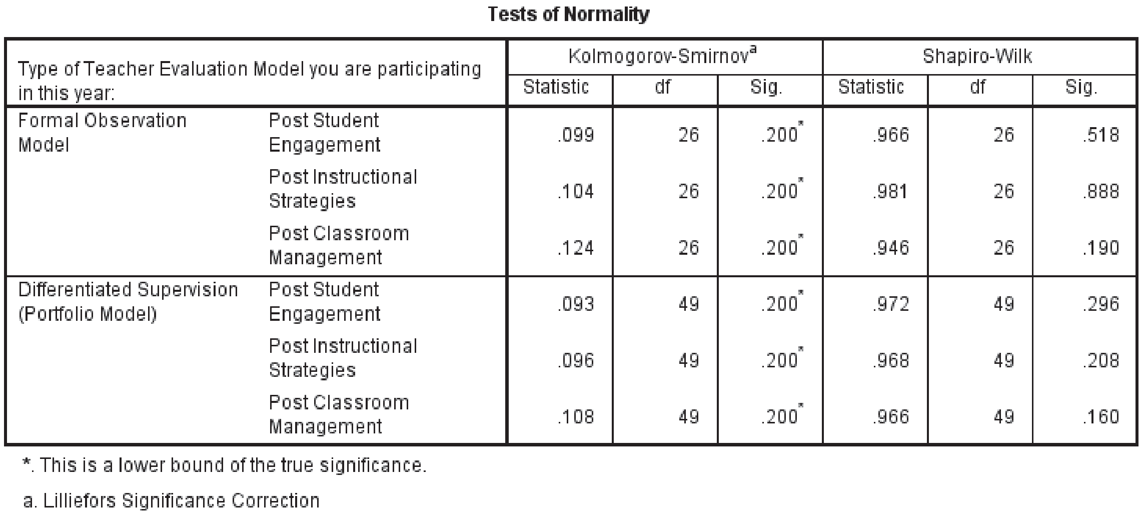


Tests of Normality

Type of Teacher Evaluation Model you are participating in this year:		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Formal Observation Model	Student Engagement	.127	31	.200*	.971	31	.556
	Instructional Strategies	.111	31	.200*	.972	31	.589
	Classroom Management	.108	31	.200*	.956	31	.222
Differentiated Supervision (Portfolio Model)	Student Engagement	.106	53	.200*	.979	53	.464
	Instructional Strategies	.070	53	.200*	.980	53	.499
	Classroom Management	.117	53	.068	.962	53	.092

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction



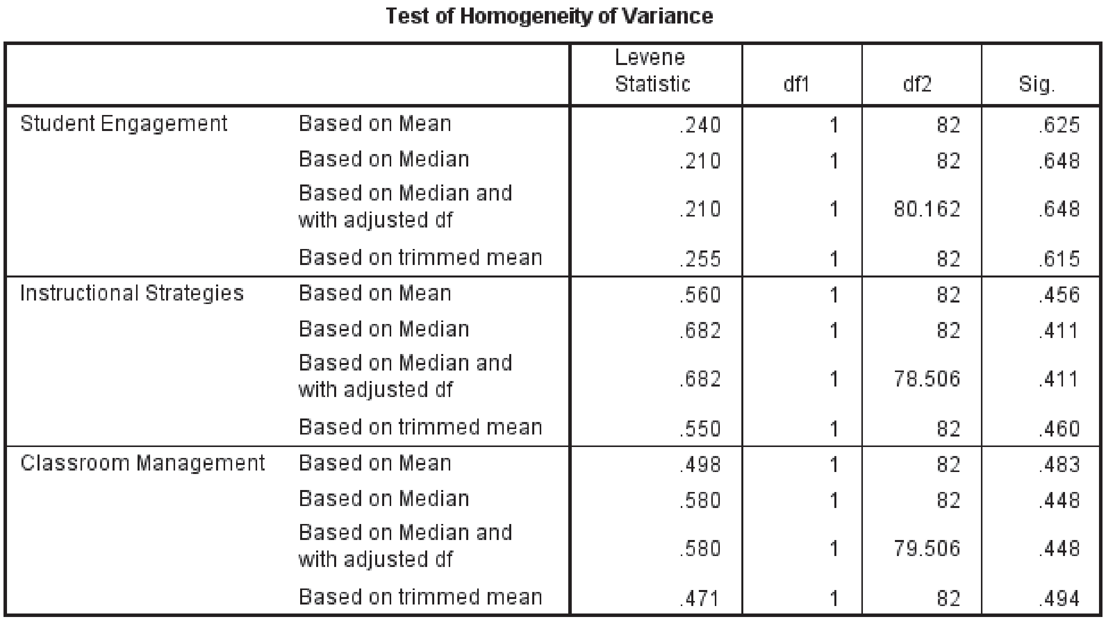
Tests of Normality

Type of Teacher Evaluation Model you are participating in this year:		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Formal Observation Model	Post Student Engagement	.099	26	.200*	.966	26	.518
	Post Instructional Strategies	.104	26	.200*	.981	26	.888
	Post Classroom Management	.124	26	.200*	.946	26	.190
Differentiated Supervision (Portfolio Model)	Post Student Engagement	.093	49	.200*	.972	49	.296
	Post Instructional Strategies	.096	49	.200*	.968	49	.208
	Post Classroom Management	.108	49	.200*	.966	49	.160

*. This is a lower bound of the true significance.

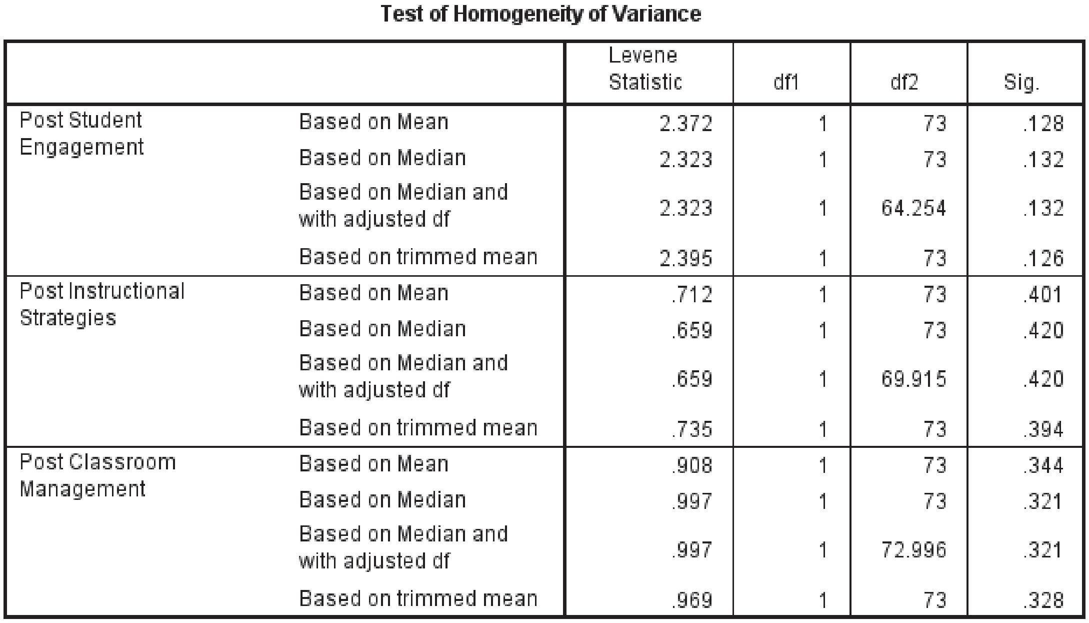
a. Lilliefors Significance Correction

Appendix S



Test of Homogeneity of Variance

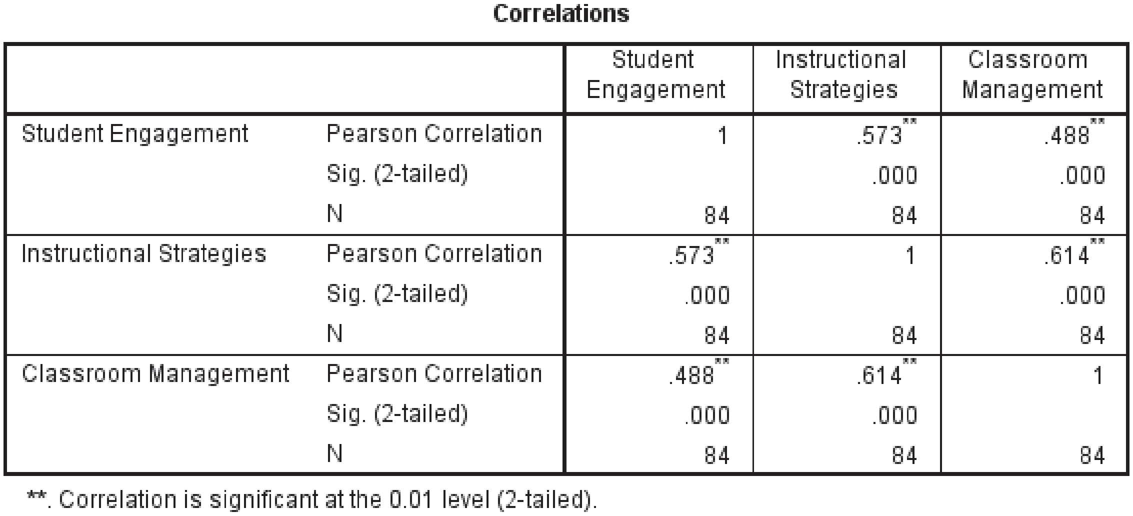
		Levene Statistic	df1	df2	Sig.
Student Engagement	Based on Mean	.240	1	82	.625
	Based on Median	.210	1	82	.648
	Based on Median and with adjusted df	.210	1	80.162	.648
	Based on trimmed mean	.255	1	82	.615
Instructional Strategies	Based on Mean	.560	1	82	.456
	Based on Median	.682	1	82	.411
	Based on Median and with adjusted df	.682	1	78.506	.411
	Based on trimmed mean	.550	1	82	.460
Classroom Management	Based on Mean	.498	1	82	.483
	Based on Median	.580	1	82	.448
	Based on Median and with adjusted df	.580	1	79.506	.448
	Based on trimmed mean	.471	1	82	.494



Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
Post Student Engagement	Based on Mean	2.372	1	73	.128
	Based on Median	2.323	1	73	.132
	Based on Median and with adjusted df	2.323	1	64.254	.132
	Based on trimmed mean	2.395	1	73	.126
Post Instructional Strategies	Based on Mean	.712	1	73	.401
	Based on Median	.659	1	73	.420
	Based on Median and with adjusted df	.659	1	69.915	.420
	Based on trimmed mean	.735	1	73	.394
Post Classroom Management	Based on Mean	.908	1	73	.344
	Based on Median	.997	1	73	.321
	Based on Median and with adjusted df	.997	1	72.996	.321
	Based on trimmed mean	.969	1	73	.328

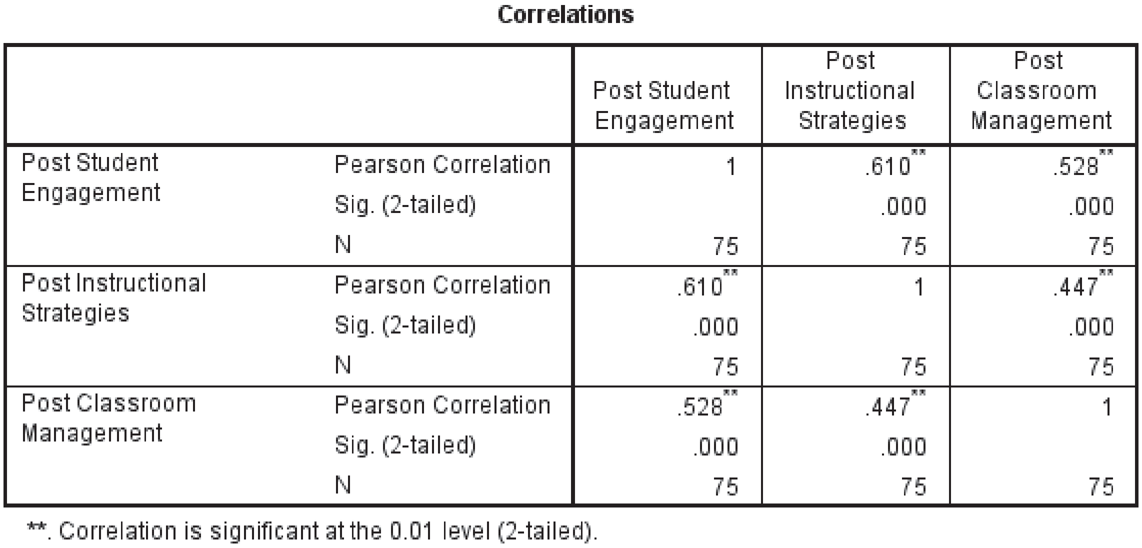
Appendix T



Correlations

		Student Engagement	Instructional Strategies	Classroom Management
Student Engagement	Pearson Correlation	1	.573**	.488**
	Sig. (2-tailed)		.000	.000
	N	84	84	84
Instructional Strategies	Pearson Correlation	.573**	1	.614**
	Sig. (2-tailed)	.000		.000
	N	84	84	84
Classroom Management	Pearson Correlation	.488**	.614**	1
	Sig. (2-tailed)	.000	.000	
	N	84	84	84

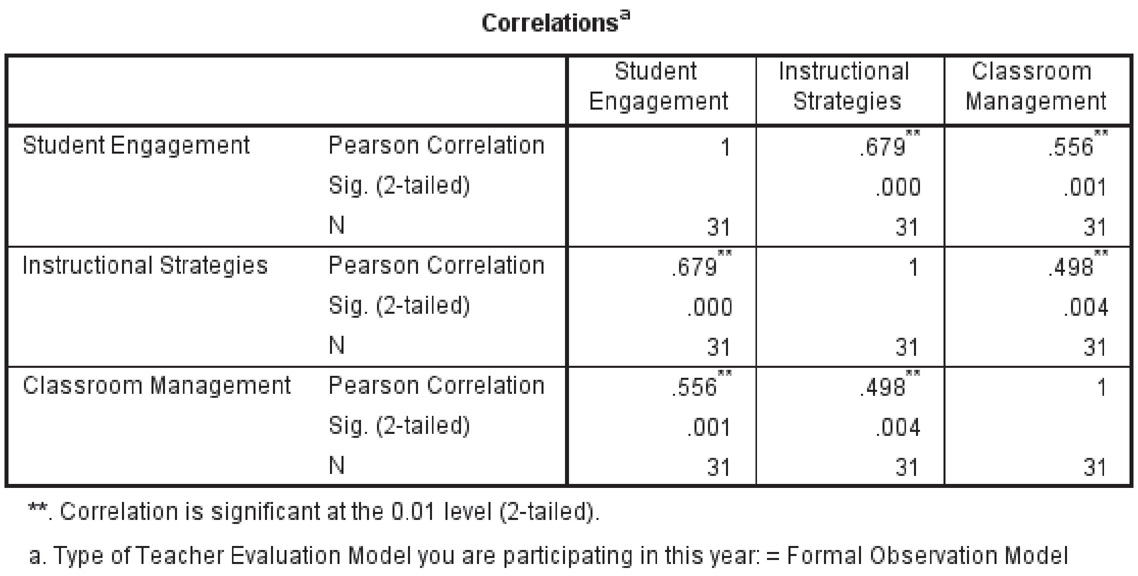
**. Correlation is significant at the 0.01 level (2-tailed).



Correlations

		Post Student Engagement	Post Instructional Strategies	Post Classroom Management
Post Student Engagement	Pearson Correlation	1	.610**	.528**
	Sig. (2-tailed)		.000	.000
	N	75	75	75
Post Instructional Strategies	Pearson Correlation	.610**	1	.447**
	Sig. (2-tailed)	.000		.000
	N	75	75	75
Post Classroom Management	Pearson Correlation	.528**	.447**	1
	Sig. (2-tailed)	.000	.000	
	N	75	75	75

**. Correlation is significant at the 0.01 level (2-tailed).

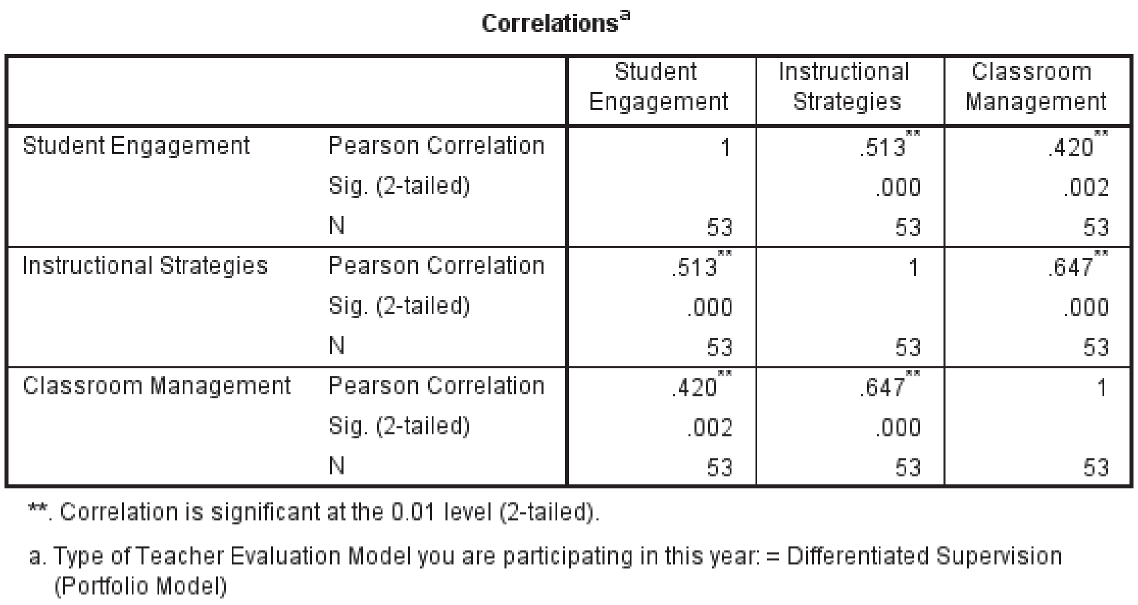


Correlations[a]

		Student Engagement	Instructional Strategies	Classroom Management
Student Engagement	Pearson Correlation	1	.679**	.556**
	Sig. (2-tailed)		.000	.001
	N	31	31	31
Instructional Strategies	Pearson Correlation	.679**	1	.498**
	Sig. (2-tailed)	.000		.004
	N	31	31	31
Classroom Management	Pearson Correlation	.556**	.498**	1
	Sig. (2-tailed)	.001	.004	
	N	31	31	31

**. Correlation is significant at the 0.01 level (2-tailed).

a. Type of Teacher Evaluation Model you are participating in this year: = Formal Observation Model

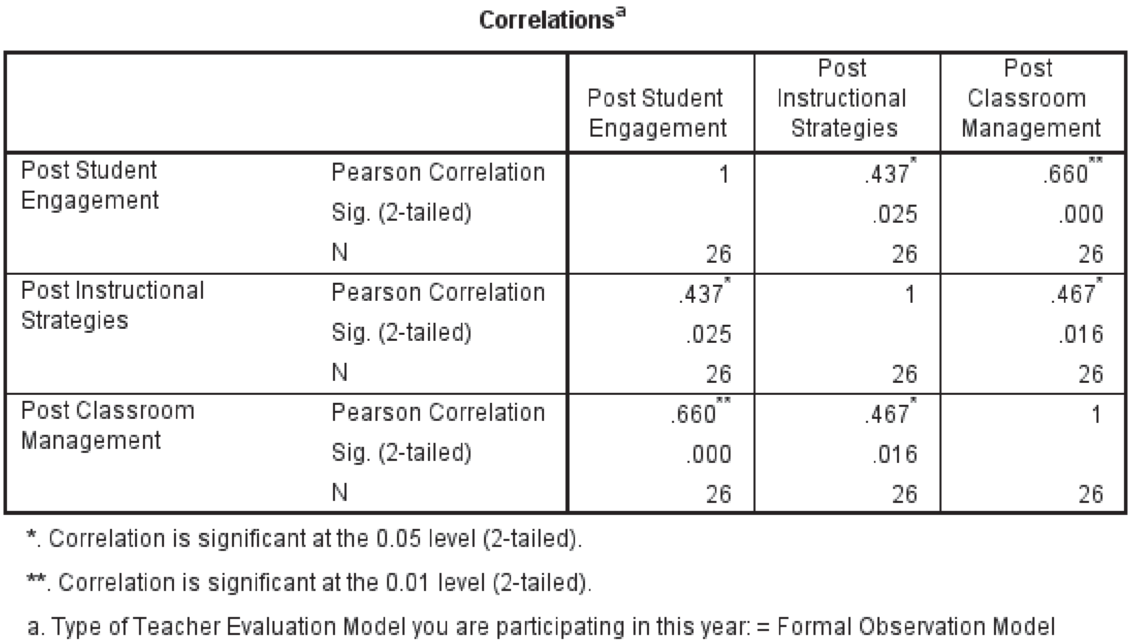


Correlations[a]

		Student Engagement	Instructional Strategies	Classroom Management
Student Engagement	Pearson Correlation	1	.513**	.420**
	Sig. (2-tailed)		.000	.002
	N	53	53	53
Instructional Strategies	Pearson Correlation	.513**	1	.647**
	Sig. (2-tailed)	.000		.000
	N	53	53	53
Classroom Management	Pearson Correlation	.420**	.647**	1
	Sig. (2-tailed)	.002	.000	
	N	53	53	53

**. Correlation is significant at the 0.01 level (2-tailed).

a. Type of Teacher Evaluation Model you are participating in this year: = Differentiated Supervision (Portfolio Model)



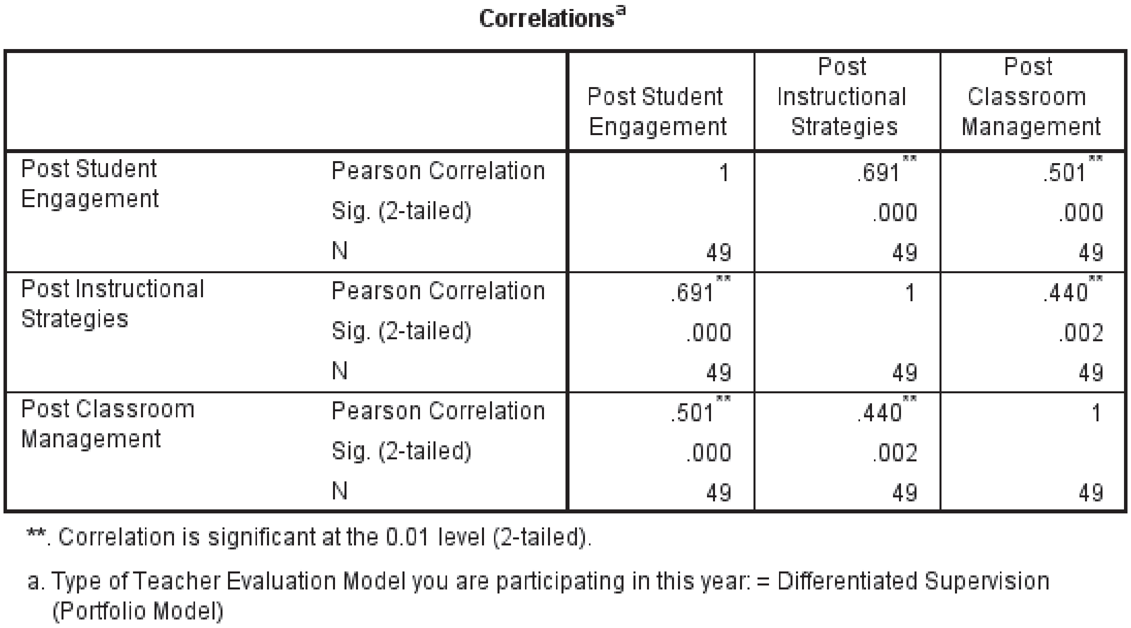
Correlations[a]

		Post Student Engagement	Post Instructional Strategies	Post Classroom Management
Post Student Engagement	Pearson Correlation	1	.437*	.660**
	Sig. (2-tailed)		.025	.000
	N	26	26	26
Post Instructional Strategies	Pearson Correlation	.437*	1	.467*
	Sig. (2-tailed)	.025		.016
	N	26	26	26
Post Classroom Management	Pearson Correlation	.660**	.467*	1
	Sig. (2-tailed)	.000	.016	
	N	26	26	26

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

a. Type of Teacher Evaluation Model you are participating in this year: = Formal Observation Model



Correlations[a]

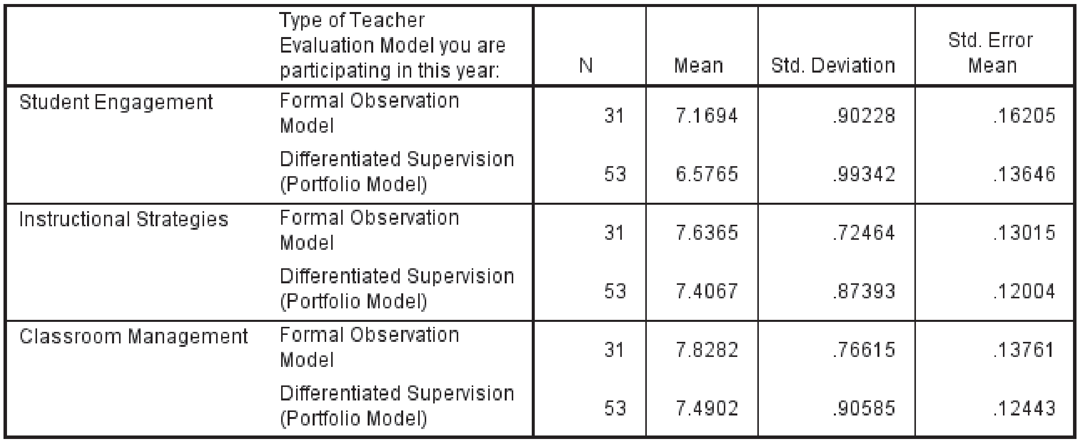
		Post Student Engagement	Post Instructional Strategies	Post Classroom Management
Post Student Engagement	Pearson Correlation	1	.691**	.501**
	Sig. (2-tailed)		.000	.000
	N	49	49	49
Post Instructional Strategies	Pearson Correlation	.691**	1	.440**
	Sig. (2-tailed)	.000		.002
	N	49	49	49
Post Classroom Management	Pearson Correlation	.501**	.440**	1
	Sig. (2-tailed)	.000	.002	
	N	49	49	49

**. Correlation is significant at the 0.01 level (2-tailed).

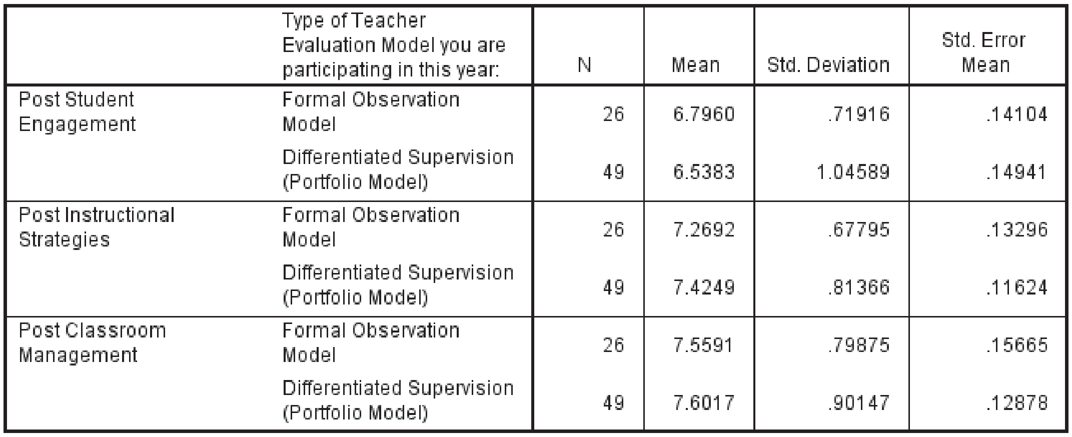
a. Type of Teacher Evaluation Model you are participating in this year: = Differentiated Supervision (Portfolio Model)

Appendix U

Group Statistics

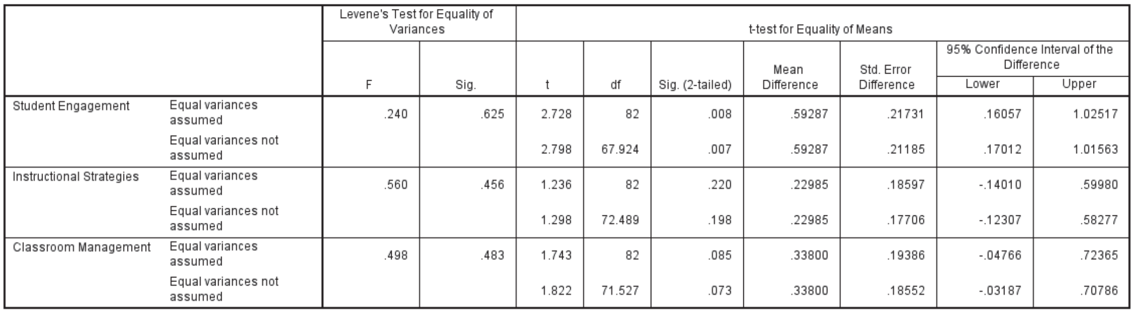


	Type of Teacher Evaluation Model you are participating in this year:	N	Mean	Std. Deviation	Std. Error Mean
Student Engagement	Formal Observation Model	31	7.1694	.90228	.16205
	Differentiated Supervision (Portfolio Model)	53	6.5765	.99342	.13646
Instructional Strategies	Formal Observation Model	31	7.6365	.72464	.13015
	Differentiated Supervision (Portfolio Model)	53	7.4067	.87393	.12004
Classroom Management	Formal Observation Model	31	7.8282	.76615	.13761
	Differentiated Supervision (Portfolio Model)	53	7.4902	.90585	.12443



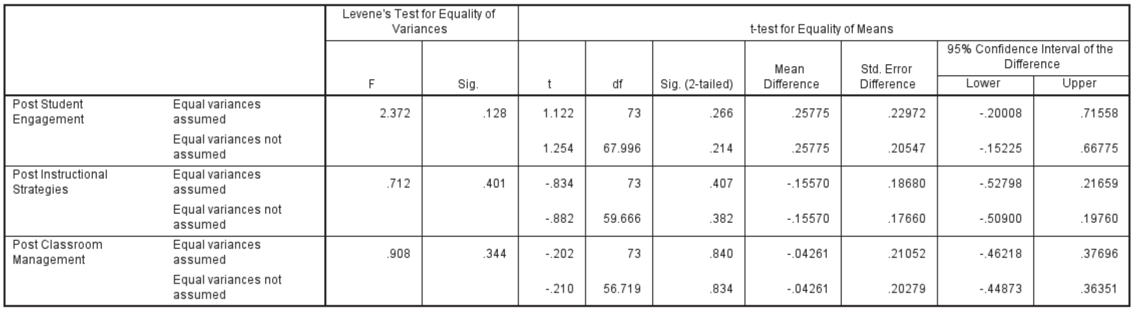
	Type of Teacher Evaluation Model you are participating in this year:	N	Mean	Std. Deviation	Std. Error Mean
Post Student Engagement	Formal Observation Model	26	6.7960	.71916	.14104
	Differentiated Supervision (Portfolio Model)	49	6.5383	1.04589	.14941
Post Instructional Strategies	Formal Observation Model	26	7.2692	.67795	.13296
	Differentiated Supervision (Portfolio Model)	49	7.4249	.81366	.11624
Post Classroom Management	Formal Observation Model	26	7.5591	.79875	.15665
	Differentiated Supervision (Portfolio Model)	49	7.6017	.90147	.12878

Independent Samples Test



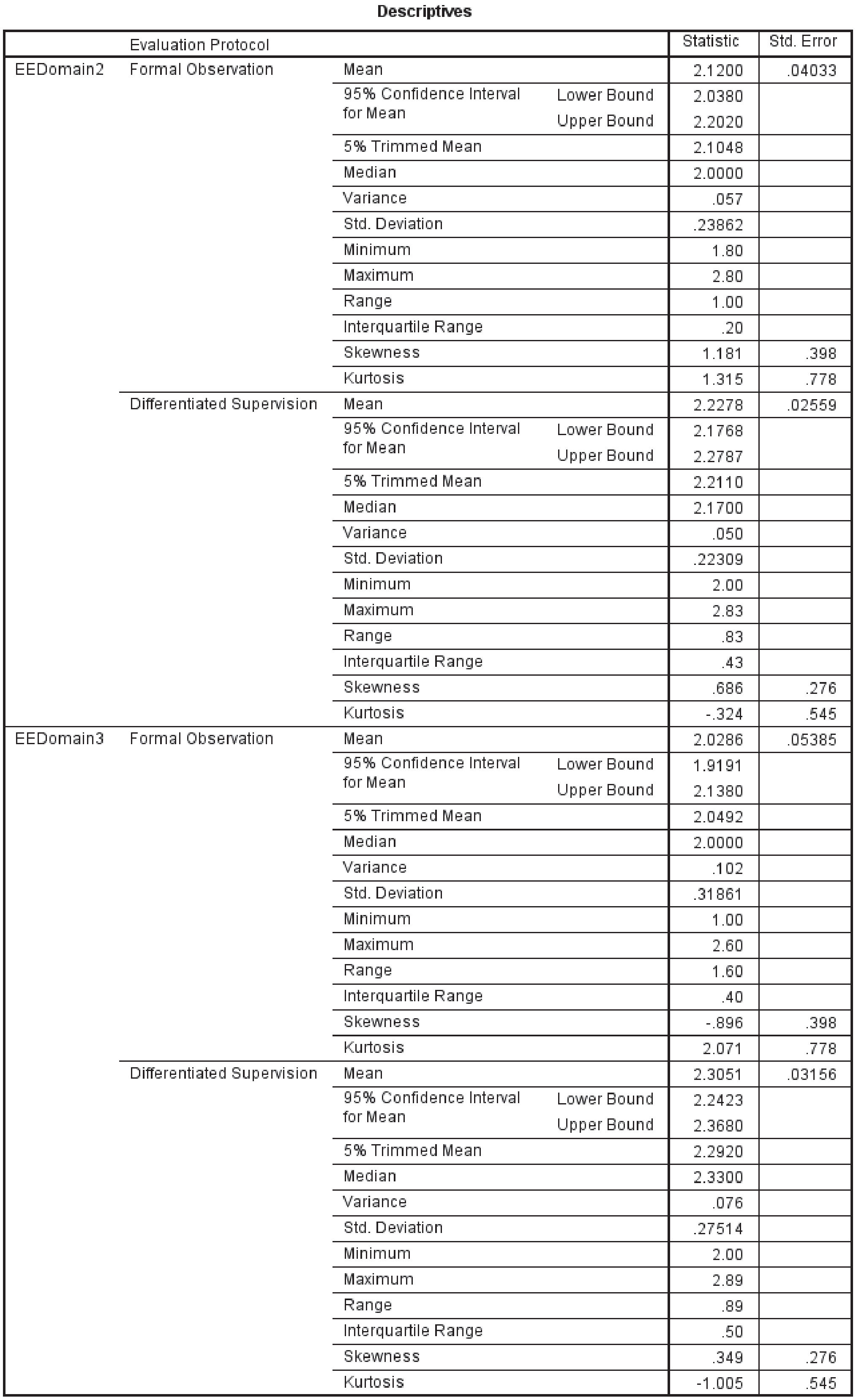
		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Student Engagement	Equal variances assumed	.240	.625	2.728	82	.008	.59287	.21731	.16057	1.02517
	Equal variances not assumed			2.798	67.924	.007	.59287	.21185	.17012	1.01563
Instructional Strategies	Equal variances assumed	.560	.456	1.236	82	.220	.22985	.18597	-.14010	.59980
	Equal variances not assumed			1.298	72.489	.198	.22985	.17706	-.12307	.58277
Classroom Management	Equal variances assumed	.498	.483	1.743	82	.085	.33800	.19386	-.04766	.72365
	Equal variances not assumed			1.822	71.527	.073	.33800	.18552	-.03187	.70786

Independent Samples Test



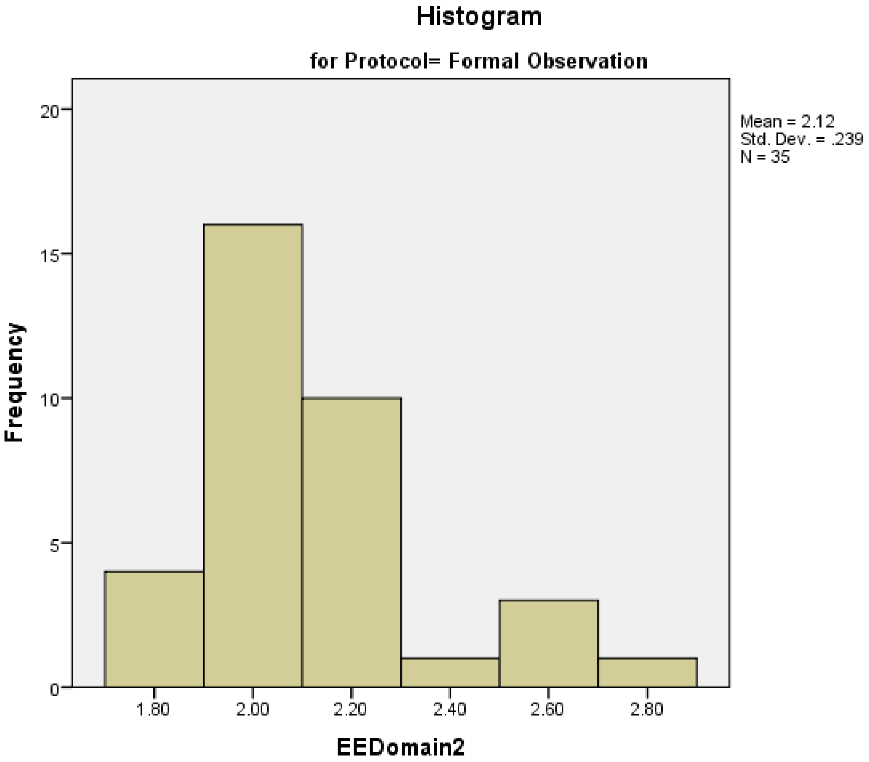
		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Post Student Engagement	Equal variances assumed	2.372	.128	1.122	73	.266	.25775	.22972	-.20008	.71558
	Equal variances not assumed			1.254	67.996	.214	.25775	.20547	-.15225	.66775
Post Instructional Strategies	Equal variances assumed	.712	.401	-.834	73	.407	-.15570	.18680	-.52798	.21659
	Equal variances not assumed			-.882	59.666	.382	-.15570	.17660	-.50900	.19760
Post Classroom Management	Equal variances assumed	.908	.344	-.202	73	.840	-.04261	.21052	-.46218	.37696
	Equal variances not assumed			-.210	56.719	.834	-.04261	.20279	-.44873	.36351

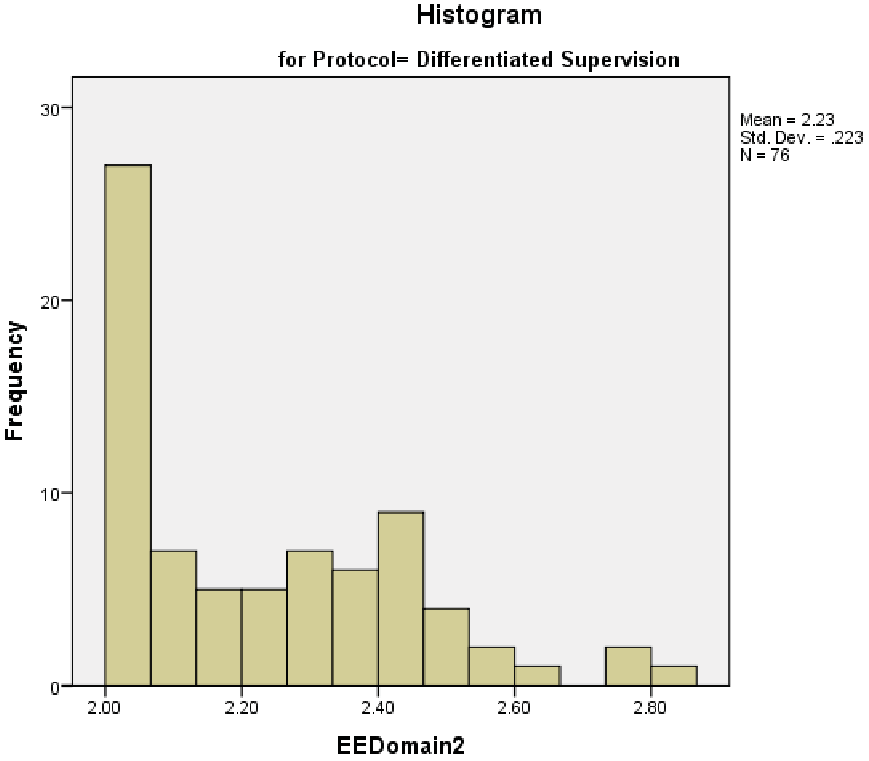
Appendix V

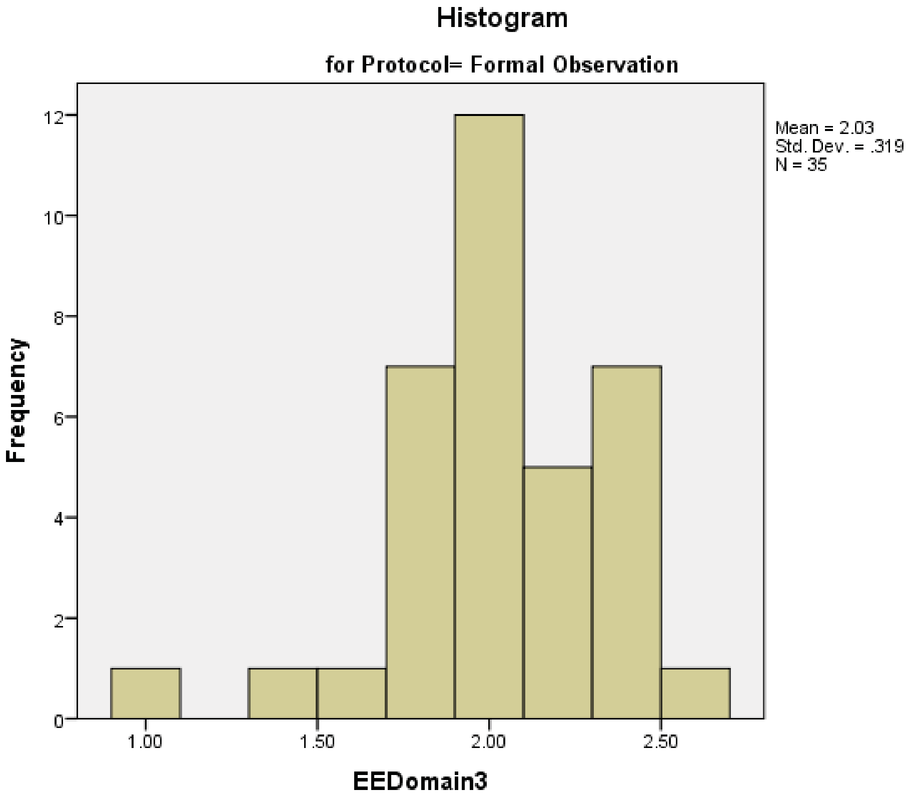


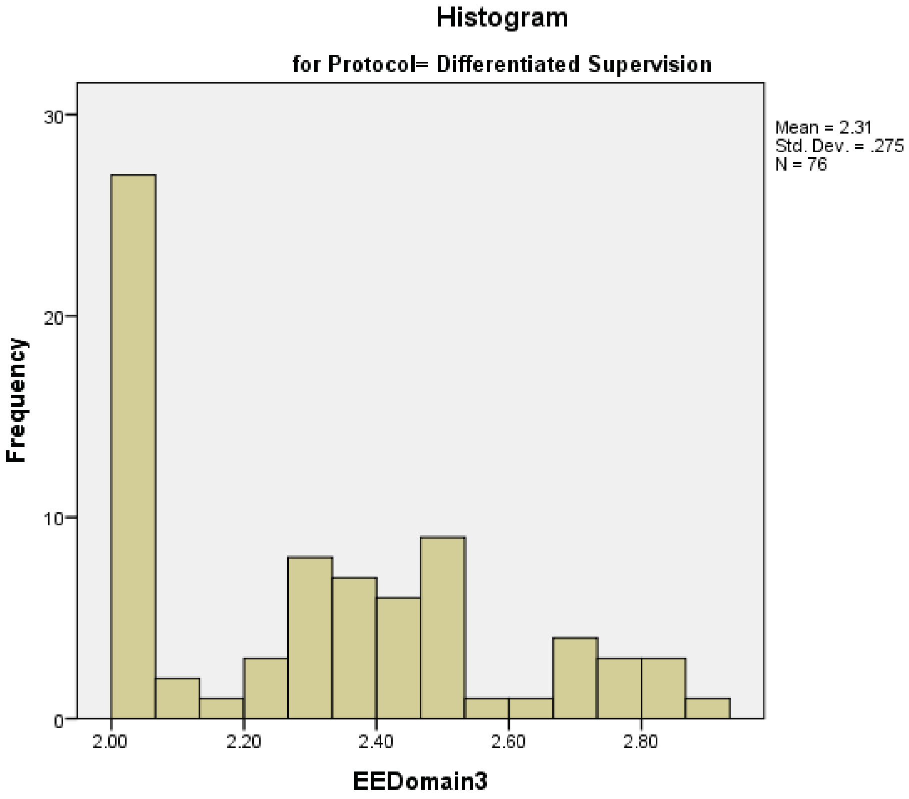
Descriptives

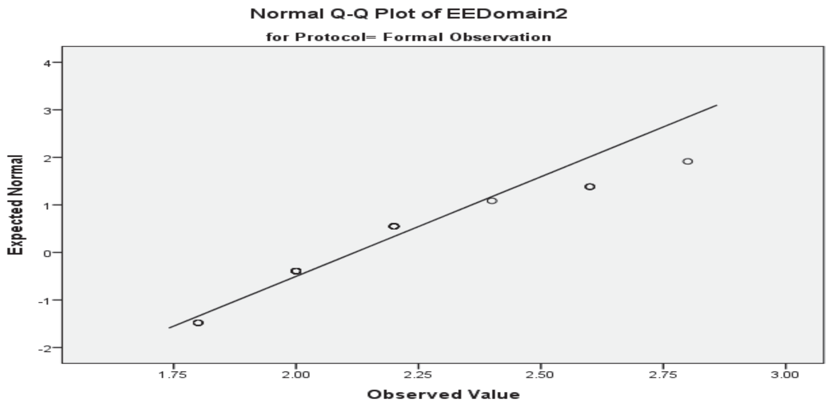
	Evaluation Protocol			Statistic	Std. Error
EEDomain2	Formal Observation	Mean		2.1200	.04033
		95% Confidence Interval for Mean	Lower Bound	2.0380	
			Upper Bound	2.2020	
		5% Trimmed Mean		2.1048	
		Median		2.0000	
		Variance		.057	
		Std. Deviation		.23862	
		Minimum		1.80	
		Maximum		2.80	
		Range		1.00	
		Interquartile Range		.20	
		Skewness		1.181	.398
		Kurtosis		1.315	.778
	Differentiated Supervision	Mean		2.2278	.02559
		95% Confidence Interval for Mean	Lower Bound	2.1768	
			Upper Bound	2.2787	
		5% Trimmed Mean		2.2110	
		Median		2.1700	
		Variance		.050	
		Std. Deviation		.22309	
		Minimum		2.00	
		Maximum		2.83	
		Range		.83	
		Interquartile Range		.43	
		Skewness		.686	.276
		Kurtosis		-.324	.545
EEDomain3	Formal Observation	Mean		2.0286	.05385
		95% Confidence Interval for Mean	Lower Bound	1.9191	
			Upper Bound	2.1380	
		5% Trimmed Mean		2.0492	
		Median		2.0000	
		Variance		.102	
		Std. Deviation		.31861	
		Minimum		1.00	
		Maximum		2.60	
		Range		1.60	
		Interquartile Range		.40	
		Skewness		-.896	.398
		Kurtosis		2.071	.778
	Differentiated Supervision	Mean		2.3051	.03156
		95% Confidence Interval for Mean	Lower Bound	2.2423	
			Upper Bound	2.3680	
		5% Trimmed Mean		2.2920	
		Median		2.3300	
		Variance		.076	
		Std. Deviation		.27514	
		Minimum		2.00	
		Maximum		2.89	
		Range		.89	
		Interquartile Range		.50	
		Skewness		.349	.276
		Kurtosis		-1.005	.545

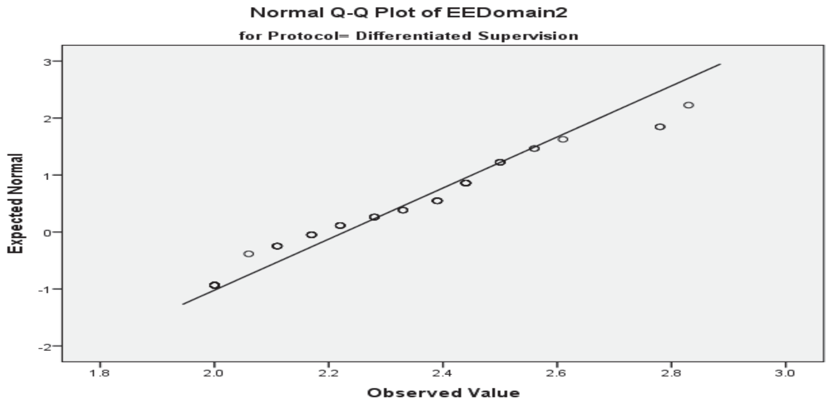

Histogram
for Protocol= Formal Observation
Frequency
20
15
10
5
0
1.80
2.00
2.20
2.40
2.60
2.80
EEDomain2
Mean = 2.12
Std. Dev. = .239
N = 35

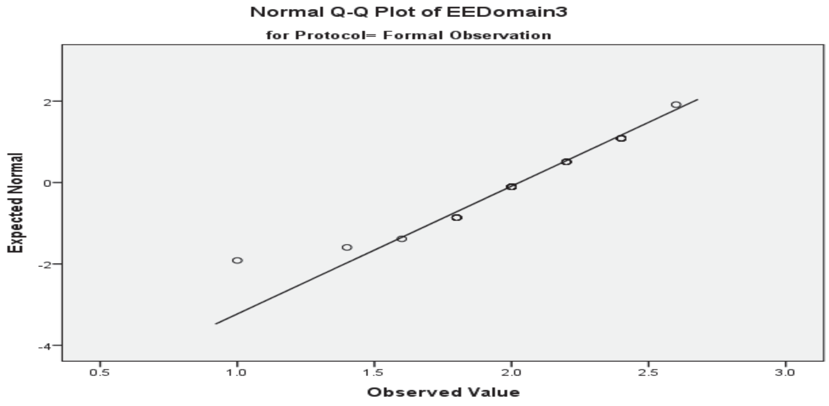

Histogram
for Protocol= Differentiated Supervision
Frequency
30
20
10
0
2.00
2.20
2.40
2.60
2.80
EEDomain2
Mean = 2.23
Std. Dev. = .223
N = 76

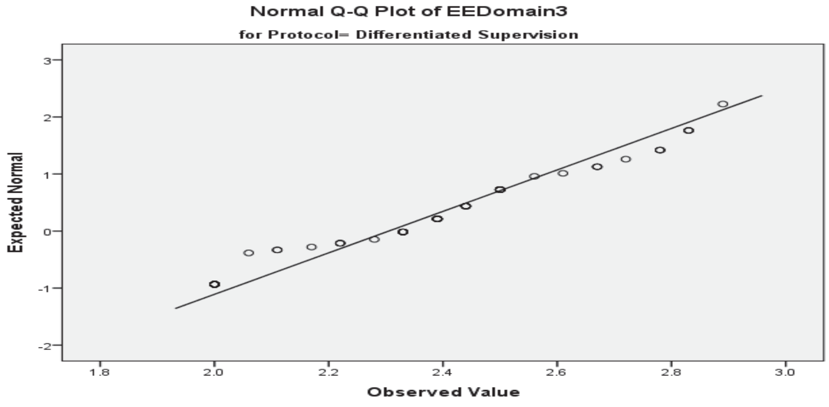

Histogram
for Protocol= Formal Observation
Mean = 2.03
Std. Dev. = .319
N = 35
Frequency
0
2
4
6
8
10
12
1.00
1.50
2.00
2.50
EEDomain3

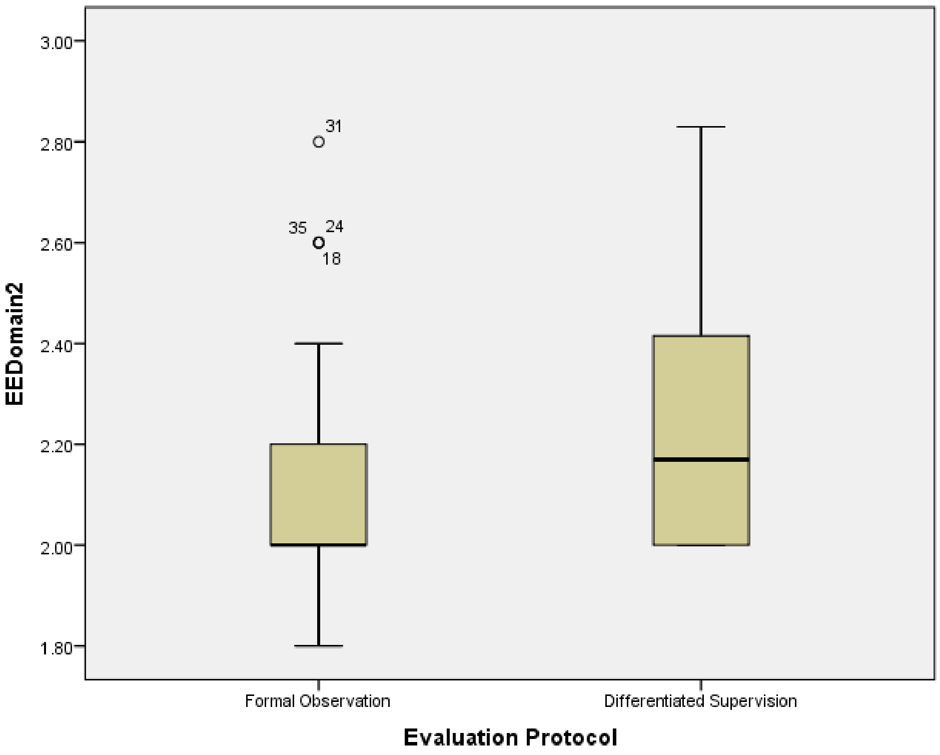

Histogram
for Protocol= Differentiated Supervision
Mean = 2.31
Std. Dev. = .275
N = 76
Frequency
0
10
20
30
2.00
2.20
2.40
2.60
2.80
EEDomain3

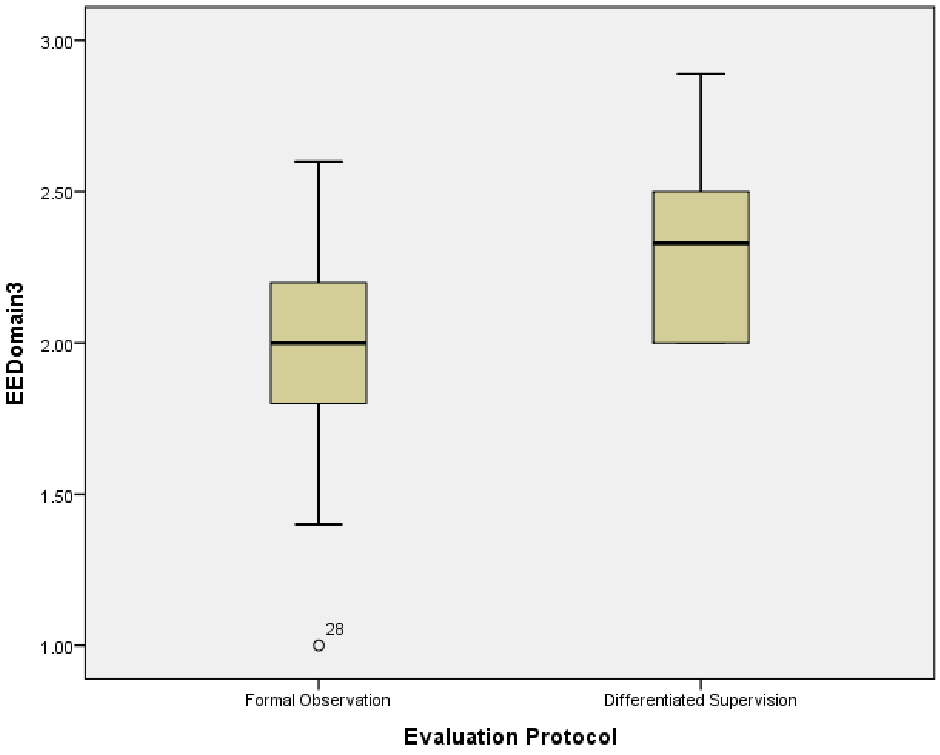

Normal Q-Q Plot of EEDomain2
for Protocol= Formal Observation
Expected Normal
4
3
2
1
0
-1
-2
1.75
2.00
2.25
2.50
2.75
3.00
Observed Value


Normal Q-Q Plot of EEDomain2
for Protocol= Differentiated Supervision
Expected Normal
3
2
1
0
-1
-2
1.8
2.0
2.2
2.4
2.6
2.8
3.0
Observed Value

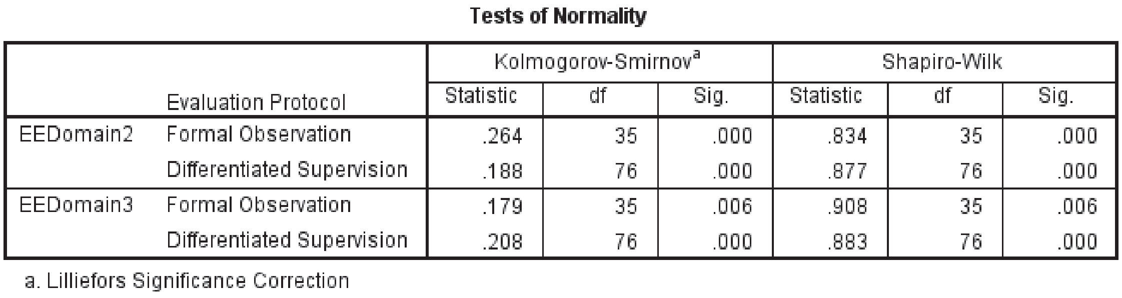

Normal Q-Q Plot of EEDomain3
for Protocol= Formal Observation
Expected Normal
2
0
-2
-4
0.5
1.0
1.5
2.0
2.5
3.0
Observed Value


Normal Q-Q Plot of EEDomain3
for Protocol= Differentiated Supervision
Expected Normal
3
2
1
0
-1
-2
1.8
2.0
2.2
2.4
2.6
2.8
3.0
Observed Value


3.00
2.80
2.60
2.40
2.20
2.00
1.80
EEDomain2
31
35
24
18
Formal Observation
Differentiated Supervision
Evaluation Protocol


3.00
2.50
2.00
1.50
1.00
EEDomain3
28
Formal Observation
Differentiated Supervision
Evaluation Protocol

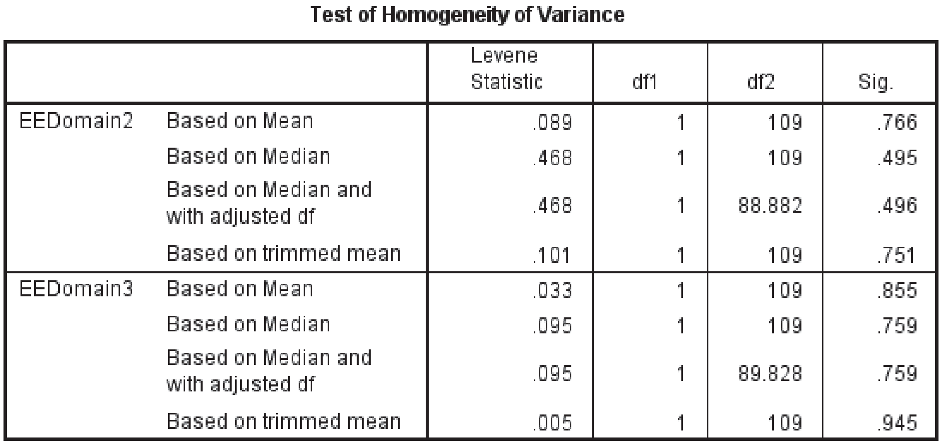
Appendix W



Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Evaluation Protocol	Statistic	df	Sig.	Statistic	df	Sig.
EEDomain2	Formal Observation	.264	35	.000	.834	35	.000
	Differentiated Supervision	.188	76	.000	.877	76	.000
EEDomain3	Formal Observation	.179	35	.006	.908	35	.006
	Differentiated Supervision	.208	76	.000	.883	76	.000

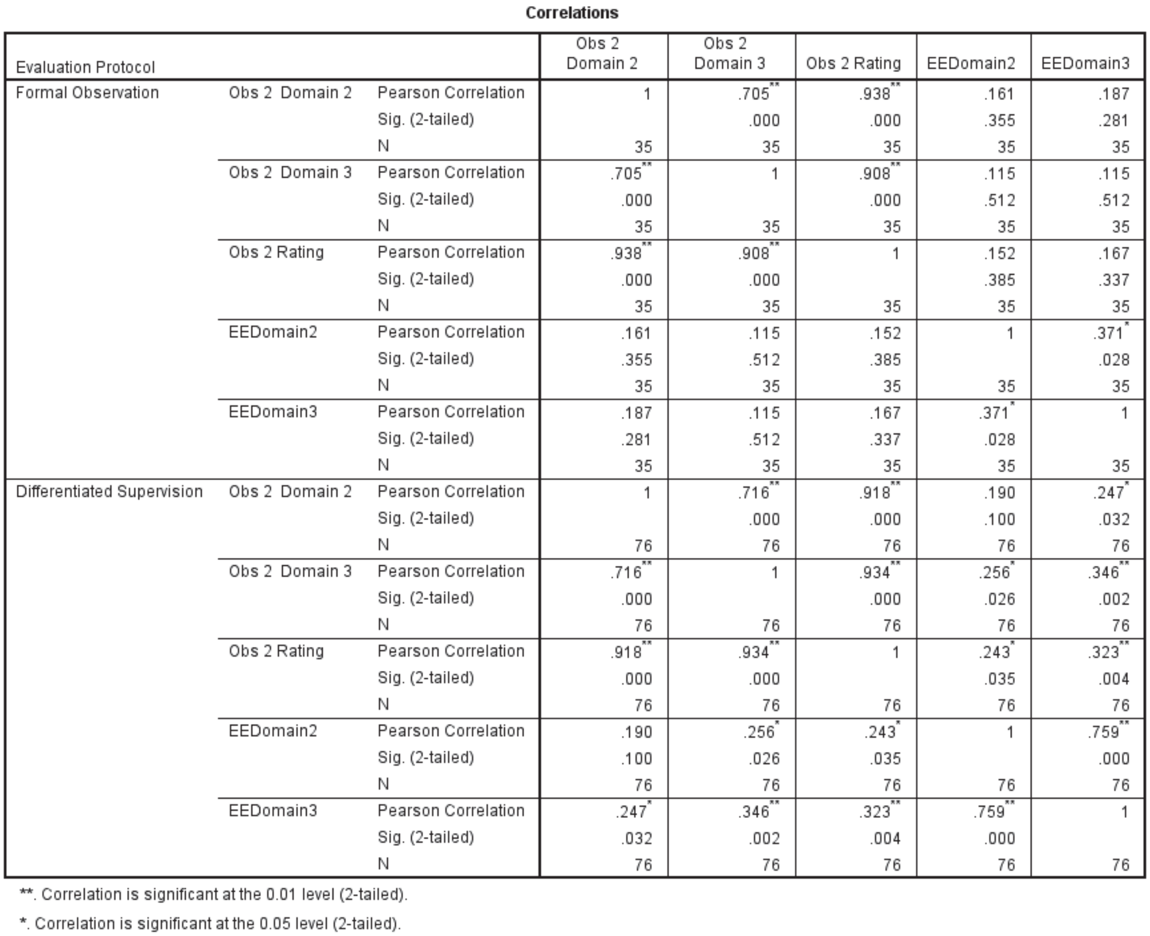
a. Lilliefors Significance Correction



Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
EEDomain2	Based on Mean	.089	1	109	.766
	Based on Median	.468	1	109	.495
	Based on Median and with adjusted df	.468	1	88.882	.496
	Based on trimmed mean	.101	1	109	.751
EEDomain3	Based on Mean	.033	1	109	.855
	Based on Median	.095	1	109	.759
	Based on Median and with adjusted df	.095	1	89.828	.759
	Based on trimmed mean	.005	1	109	.945

Appendix X



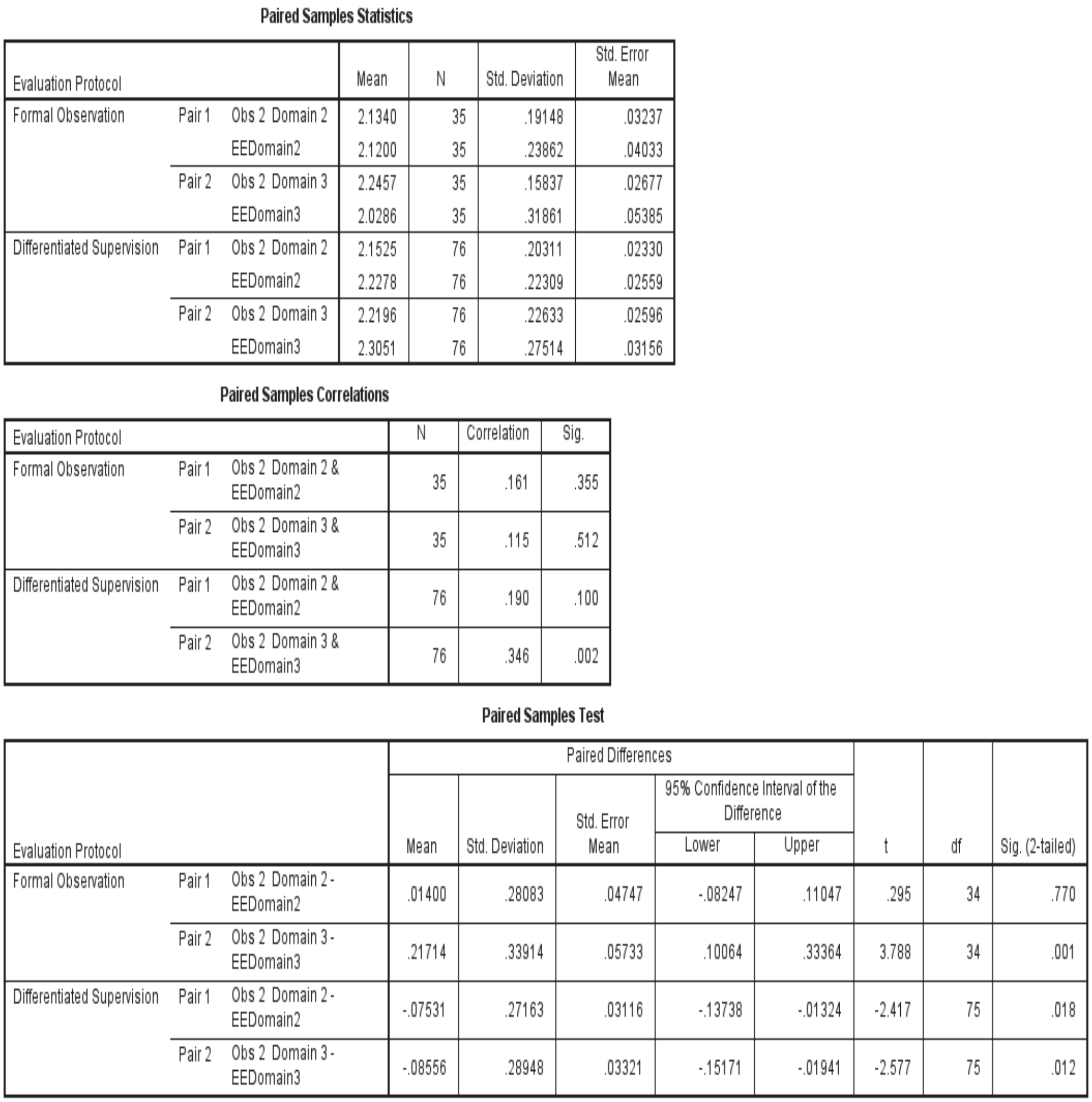
Correlations

Evaluation Protocol			Obs 2 Domain 2	Obs 2 Domain 3	Obs 2 Rating	EEDomain2	EEDomain3
Formal Observation	Obs 2 Domain 2	Pearson Correlation	1	.705**	.938**	.161	.187
		Sig. (2-tailed)		.000	.000	.355	.281
		N	35	35	35	35	35
	Obs 2 Domain 3	Pearson Correlation	.705**	1	.908**	.115	.115
		Sig. (2-tailed)	.000		.000	.512	.512
		N	35	35	35	35	35
	Obs 2 Rating	Pearson Correlation	.938**	.908**	1	.152	.167
		Sig. (2-tailed)	.000	.000		.385	.337
		N	35	35	35	35	35
	EEDomain2	Pearson Correlation	.161	.115	.152	1	.371*
		Sig. (2-tailed)	.355	.512	.385		.028
		N	35	35	35	35	35
	EEDomain3	Pearson Correlation	.187	.115	.167	.371*	1
		Sig. (2-tailed)	.281	.512	.337	.028	
		N	35	35	35	35	35
Differentiated Supervision	Obs 2 Domain 2	Pearson Correlation	1	.716**	.918**	.190	.247*
		Sig. (2-tailed)		.000	.000	.100	.032
		N	76	76	76	76	76
	Obs 2 Domain 3	Pearson Correlation	.716**	1	.934**	.256*	.346**
		Sig. (2-tailed)	.000		.000	.026	.002
		N	76	76	76	76	76
	Obs 2 Rating	Pearson Correlation	.918**	.934**	1	.243*	.323**
		Sig. (2-tailed)	.000	.000		.035	.004
		N	76	76	76	76	76
	EEDomain2	Pearson Correlation	.190	.256*	.243*	1	.759**
		Sig. (2-tailed)	.100	.026	.035		.000
		N	76	76	76	76	76
	EEDomain3	Pearson Correlation	.247*	.346**	.323**	.759**	1
		Sig. (2-tailed)	.032	.002	.004	.000	
		N	76	76	76	76	76

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Appendix Y



Paired Samples Statistics

Evaluation Protocol			Mean	N	Std. Deviation	Std. Error Mean
Formal Observation	Pair 1	Obs 2 Domain 2	2.1340	35	.19148	.03237
		EEDomain2	2.1200	35	.23862	.04033
	Pair 2	Obs 2 Domain 3	2.2457	35	.15837	.02677
		EEDomain3	2.0286	35	.31861	.05385
Differentiated Supervision	Pair 1	Obs 2 Domain 2	2.1525	76	.20311	.02330
		EEDomain2	2.2278	76	.22309	.02559
	Pair 2	Obs 2 Domain 3	2.2196	76	.22633	.02596
		EEDomain3	2.3051	76	.27514	.03156

Paired Samples Correlations

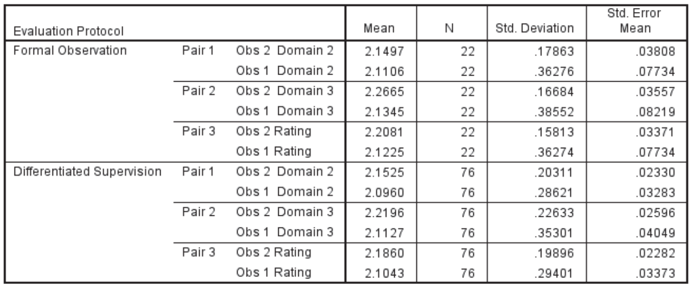
Evaluation Protocol			N	Correlation	Sig.
Formal Observation	Pair 1	Obs 2 Domain 2 & EEDomain2	35	.161	.355
	Pair 2	Obs 2 Domain 3 & EEDomain3	35	.115	.512
Differentiated Supervision	Pair 1	Obs 2 Domain 2 & EEDomain2	76	.190	.100
	Pair 2	Obs 2 Domain 3 & EEDomain3	76	.346	.002

Paired Samples Test

			Paired Differences							
						95% Confidence Interval of the Difference				
Evaluation Protocol			Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Formal Observation	Pair 1	Obs 2 Domain 2 - EEDomain2	.01400	.28083	.04747	-.08247	.11047	.295	34	.770
	Pair 2	Obs 2 Domain 3 - EEDomain3	.21714	.33914	.05733	.10064	.33364	3.788	34	.001
Differentiated Supervision	Pair 1	Obs 2 Domain 2 - EEDomain2	-.07531	.27163	.03116	-.13738	-.01324	-2.417	75	.018
	Pair 2	Obs 2 Domain 3 - EEDomain3	-.08556	.28948	.03321	-.15171	-.01941	-2.577	75	.012

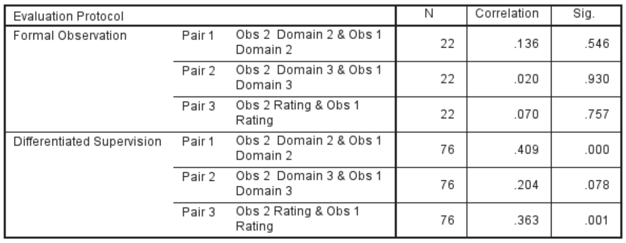
Appendix Z

Paired Samples Statistics



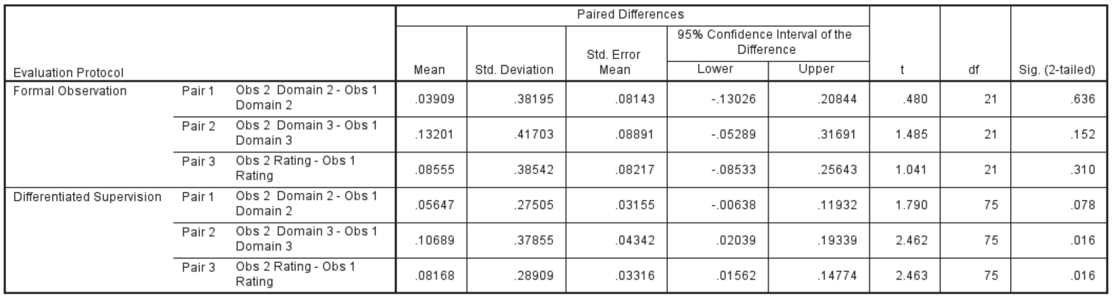
Evaluation Protocol			Mean	N	Std. Deviation	Std. Error Mean
Formal Observation	Pair 1	Obs 2 Domain 2	2.1497	22	.17863	.03808
		Obs 1 Domain 2	2.1106	22	.36276	.07734
	Pair 2	Obs 2 Domain 3	2.2665	22	.16684	.03557
		Obs 1 Domain 3	2.1345	22	.38552	.08219
	Pair 3	Obs 2 Rating	2.2081	22	.15813	.03371
		Obs 1 Rating	2.1225	22	.36274	.07734
Differentiated Supervision	Pair 1	Obs 2 Domain 2	2.1525	76	.20311	.02330
		Obs 1 Domain 2	2.0960	76	.28621	.03283
	Pair 2	Obs 2 Domain 3	2.2196	76	.22633	.02596
		Obs 1 Domain 3	2.1127	76	.35301	.04049
	Pair 3	Obs 2 Rating	2.1860	76	.19896	.02282
		Obs 1 Rating	2.1043	76	.29401	.03373

Paired Samples Correlations



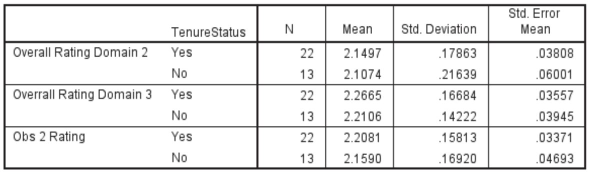
Evaluation Protocol			N	Correlation	Sig.
Formal Observation	Pair 1	Obs 2 Domain 2 & Obs 1 Domain 2	22	.136	.546
	Pair 2	Obs 2 Domain 3 & Obs 1 Domain 3	22	.020	.930
	Pair 3	Obs 2 Rating & Obs 1 Rating	22	.070	.757
Differentiated Supervision	Pair 1	Obs 2 Domain 2 & Obs 1 Domain 2	76	.409	.000
	Pair 2	Obs 2 Domain 3 & Obs 1 Domain 3	76	.204	.078
	Pair 3	Obs 2 Rating & Obs 1 Rating	76	.363	.001

Paired Samples Test



			Paired Differences							
						95% Confidence Interval of the Difference				
Evaluation Protocol			Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Formal Observation	Pair 1	Obs 2 Domain 2 - Obs 1 Domain 2	.03909	.38195	.08143	-.13026	.20844	.480	21	.636
	Pair 2	Obs 2 Domain 3 - Obs 1 Domain 3	.13201	.41703	.08891	-.05289	.31691	1.485	21	.152
	Pair 3	Obs 2 Rating - Obs 1 Rating	.08555	.38542	.08217	-.08533	.25643	1.041	21	.310
Differentiated Supervision	Pair 1	Obs 2 Domain 2 - Obs 1 Domain 2	.05647	.27505	.03155	-.00638	.11932	1.790	75	.078
	Pair 2	Obs 2 Domain 3 - Obs 1 Domain 3	.10689	.37855	.04342	.02039	.19339	2.462	75	.016
	Pair 3	Obs 2 Rating - Obs 1 Rating	.08168	.28909	.03316	.01562	.14774	2.463	75	.016

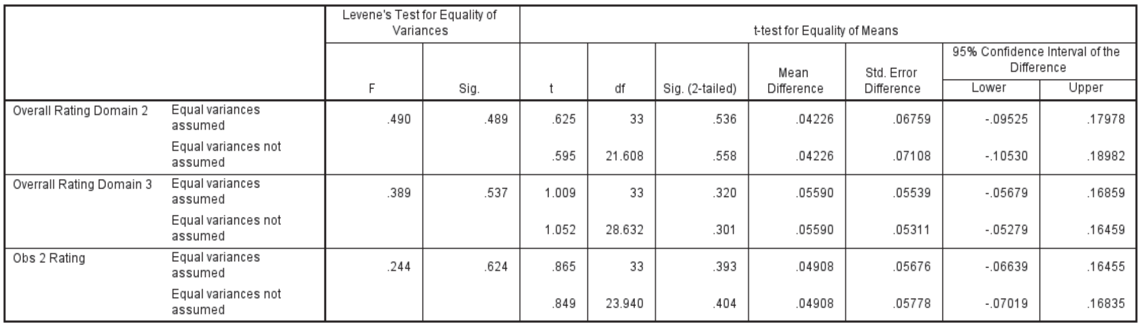
Group Statistics[a]



	TenureStatus	N	Mean	Std. Deviation	Std. Error Mean
Overall Rating Domain 2	Yes	22	2.1497	.17863	.03808
	No	13	2.1074	.21639	.06001
Overall Rating Domain 3	Yes	22	2.2665	.16684	.03557
	No	13	2.2106	.14222	.03945
Obs 2 Rating	Yes	22	2.2081	.15813	.03371
	No	13	2.1590	.16920	.04693

a. Evaluation Protocol = Formal Observation

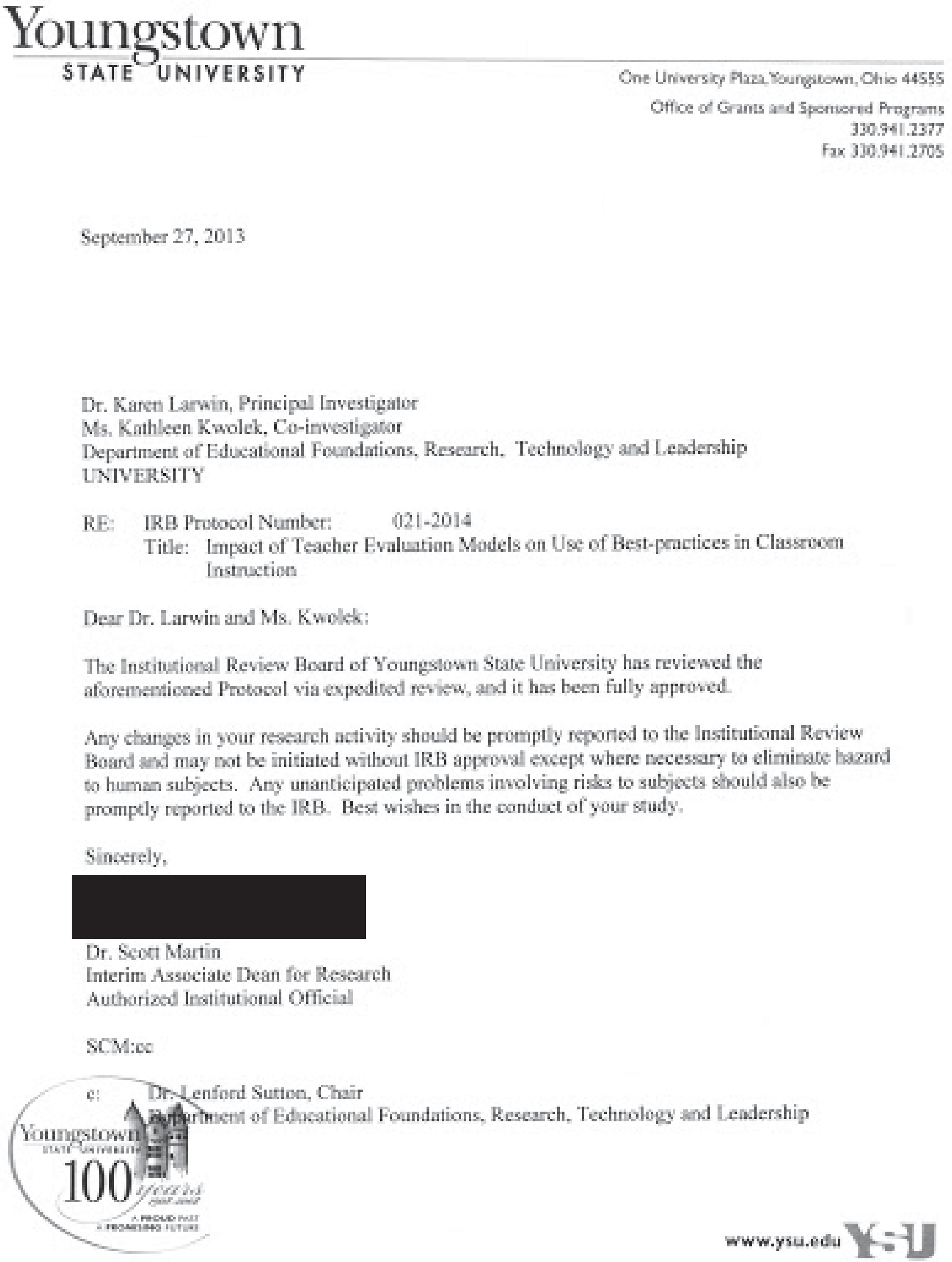
Independent Samples Test[a]



		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Overall Rating Domain 2	Equal variances assumed	.490	.489	.625	33	.536	.04226	.06759	-.09525	.17978
	Equal variances not assumed			.595	21.608	.558	.04226	.07108	-.10530	.18982
Overall Rating Domain 3	Equal variances assumed	.389	.537	1.009	33	.320	.05590	.05539	-.05679	.16859
	Equal variances not assumed			1.052	28.632	.301	.05590	.05311	-.05279	.16459
Obs 2 Rating	Equal variances assumed	.244	.624	.865	33	.393	.04908	.05676	-.06639	.16455
	Equal variances not assumed			.849	23.940	.404	.04908	.05778	-.07019	.16835

a. Evaluation Protocol = Formal Observation

Appendix AA



Youngstown
STATE UNIVERSITY

One University Plaza, Youngstown, Ohio 44555
Office of Grants and Sponsored Programs
330.941.2377
Fax 330.941.2705

September 27, 2013

Dr. Karen Larwin, Principal Investigator
Ms. Kathleen Kwolek, Co-investigator
Department of Educational Foundations, Research, Technology and Leadership
UNIVERSITY

RE: IRB Protocol Number: 021-2014
Title: Impact of Teacher Evaluation Models on Use of Best-practices in Classroom Instruction

Dear Dr. Larwin and Ms. Kwolek:

The Institutional Review Board of Youngstown State University has reviewed the aforementioned Protocol via expedited review, and it has been fully approved.

Any changes in your research activity should be promptly reported to the Institutional Review Board and may not be initiated without IRB approval except where necessary to eliminate hazard to human subjects. Any unanticipated problems involving risks to subjects should also be promptly reported to the IRB. Best wishes in the conduct of your study.

Sincerely,

[redacted]

Dr. Scott Martin
Interim Associate Dean for Research
Authorized Institutional Official

SCM:cc

c: Dr. Lenford Sutton, Chair
Department of Educational Foundations, Research, Technology and Leadership

www.ysu.edu YSU