

IMPROVING EDUCATOR EVALUATION IN NORTH CAROLINA

A disquisition presented to the faculty of the Graduate School of
Western Carolina University in partial fulfillment of the
Requirements for the degree of Doctor of Education

By

Kimberly Simmons

Director: Dr. Kathleen Topolka-Jorissen
Associate Professor
Department of Human Services

Committee Members:
Dr. Robert Crow
Dr. Lynne Johnson
Dr. John Sherlock

March 2016

© 2016 by Kimberly Simmons

TABLE OF CONTENTS

	Page
List of Tables.....	4
List of Figures	5
Abstract.....	6
Introduction to the Project	8
Problem Identification.....	8
Educator Evaluation Systems.....	8
Framing the Problem.....	9
Research Evidence of the Problem	14
Evaluation Training	17
Conceptual Framework	18
Desired State.....	19
Setting and Context of the Improvement Project	21
Historical and Current Initiatives.....	21
Recent Initiatives and the Call for Improvement in Teacher Evaluation	23
Intervention Design	29
Intervention Design Team.....	29
Intervention Process.....	31
Overview of the Improvement Model	31
Data Sources and Uses	33
Intervention Procedures.....	34
Data Analysis.....	35
Implementation Plan Including Projected SMART Goals That Spanned the Project	37
List of Supporting Artifacts.....	37
Implementation and Results.....	39
Intervention Cycles	40
Pre-Cycle Scoring Study 1	40
Identification of areas of agreement among raters.....	40
Identification of participation related to performance	42
Intervention Cycle 1: Webinar	44
Strategies for increasing personal accuracy	45
Scripting quality evidence	46
Developing a personal plan	47
Monthly monitoring of participation and performance.....	47
Addressing low participation in the online training.....	49
Intervention Cycle 2: Webinar	49
Follow-up support for webinar participants	51
Participant frustration feedback.....	52
Intervention Cycle 3: Increase Participation.....	53
Data from SS2	55
Improvement in Participant Performance	57

Improvement in Scoring Accuracy.....	58
Identification of difficult to rate elements.....	59
Impact of Number of Lessons Completed	61
Summary of Improvement Evidence	63
Impact, Discussion, and Steps.....	64
Continuation of Intervention for North Carolina.....	64
Recommended Changes to the Intervention Process	64
Impact of Collaborative Participation.....	66
Additional State Recommendations	69
Policy Considerations	69
Final Thoughts	70
References.....	71
Appendix: Supporting Artifacts	76

LIST OF TABLES

Table	Page
1. 2014–15 School Accountability Growth.....	23
2. Implementation Plan	37
3. Elements Scored Incorrectly Most Often 2/19/15	48
4. 2014–15 OCT Participation as of March 23, 2015.....	52

LIST OF FIGURES

Figure	Page
1. Flat organization chart	13
2. Conceptual framework.....	19
3. PDSA Cycle and Model for Improvement, 1991, 1994.....	33
4. Intervention framework.....	39
5. Participant agreement with target scores by domain	43
6. Scoring Study 1 identified elements for further study.....	44
7. Analyzing disagreement: Sample scoring distribution for ST3d	46
8. Scoring study comparison: Overall agreement.....	58
9. Scoring study comparison: Agreement by element	60
10. Scoring Study 1 & Scoring Study 2 group performance	60
11. Regression model: Effect of lessons on target agreement	62
12. Regression model: Effect of lessons on target discrepancy	63

ABSTRACT

IMPROVING EDUCATOR EVALUATION IN NORTH CAROLINA

Kimberly Simmons, Ed.D.

Western Carolina University (January 2016)

Director: Dr. Kathleen Topolka-Jorissen

The purpose of this project was to improve ways the North Carolina Department of Public Instruction can develop and implement a process for improving rater agreement performance on Standards 1 through 5 of the North Carolina Professional Teacher Standards. Specifically, I used improvement science methods and the progress of the 360 participants in a pilot of the state's web-based Observation Calibration Training (OCT) system over a seven-month period. Data analysis included a pretest and posttest to determine improvement in rater agreement from the participation of the pilot program OCT. Additional data analysis followed each of the three interventions implemented during the course of the OCT to improve rater agreement and inform project modifications and next steps, with the goal of improving the participants' teacher evaluation competence. A plan for periodic assessment of change and analysis of progress toward improvement included monthly performance data reports, including participation and performance, to ensure that the OCT pilot was progressing for participants and to identify the elements, from the Professional Teacher Standards, that participants were scoring correctly and incorrectly. Based on these data, I developed and facilitated webinar interventions and supporting resources to address identified rating

performance issues. The goal was to use the results of this project to inform phase 2 of the OCT statewide training plan in ways that will increase the likelihood of participants' improvement in rating teaching behaviors.

CHAPTER ONE: INTRODUCTION TO THE PROJECT

Problem Identification

Educator Evaluation Systems

Most current systems for evaluating educators are ineffective and have little impact on educator growth and student learning and development (Fetter, 2013).

Research on educator evaluation finds that evaluations and judgments of effectiveness are often focused on superficial measures that are not linked to students' learning and outcomes (Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2011). It also shows that evaluations are rarely part of an effective system of professional growth and development (Darling-Hammond et al., 2011). Although most systems of educator evaluation are ineffective, some systems have shown success. States and districts across the nation are currently instituting holistic and diverse systems of educator evaluation (Darling-Hammond et al., 2011). In North Carolina developing an effective educator evaluation system has become a priority among policymakers and practitioners.

Classroom observations are a significant part of the evaluation process. Educators and policymakers agree that the classroom observation process has the potential to improve teacher skills and student achievement (Gandha & Baxter, 2015). The importance of observation is described in terms of its processes and outcomes by scholars who say,

Classroom observation is a powerful component of teacher evaluation systems. It measures instructional practice, provides clarification on what effective teaching looks like and gives teachers the concrete and actionable feedback they need to improve teaching practice. (Gandha & Baxter, 2015, p. 3)

However, in order to achieve these outcomes, observations must be fair and reliable to have influence on instructional improvement.

Evaluator/observer training to ensure a valid measure of teacher quality is difficult and presents complex challenges. As scholars point out, “The ongoing challenge for many states is developing an accurate understanding of different levels of teaching quality that is shared by all educators. What is considered distinguished, proficient or unsatisfactory” (Gandha & Baxter, 2015, p. 7). Effective training for evaluators and educator support for improvement is necessary to develop skilled evaluators who can provide relevant feedback to promote teacher growth and improvement. Many states, including North Carolina have developed rubrics for evaluators to use when observing a teacher. However, North Carolina educators have clearly communicated that they want more specific resources, including a checklist, outlining what they should see when observing a teacher. Additional tools, resources, and training are needed to support educators’ understanding of effective teaching practices when conducting an observation (Gandha & Baxter, 2015). Complicating this goal further is the subjective reality of teacher observation, in which what observers already know about a teacher prior to an observation influences their judgments (Gandha & Baxter, 2015, p. 10). The complexity of the teacher evaluation process presents a range of problems for policymakers and leaders committed to using evaluation to improve teacher quality.

Framing the Problem

In the last few years, several models of teacher evaluation and observation have been developed in order to help educators make decisions about the effectiveness of teaching and learning in schools and to seek their improvement. These models include

using ideas from other professions such as medical rounds and review boards, using student test or growth scores, and creating comprehensive or multifaceted systems (Fetter, 2013). It is clear from the review of these different models of evaluation and assessment of educators that the new North Carolina model is attempting to combine some of the best ideas from the last decade into a holistic model of evaluation that promotes the growth and development of teachers. With the adoption of College and Career Readiness standards for students in North Carolina, even greater emphasis is being placed on the quality of the teachers. For most policy makers, district leaders, and scholars one essential determinant of quality is teacher evaluation (Simmons & Mullins, 2013). The challenge of ensuring that evaluators rate teachers accurately must be addressed if the goal of providing quality teachers for all children is to be realized.

North Carolina's PK-12 teachers deserve to be assessed by skilled evaluators who can provide clear, specific, constructive feedback to support both their daily classroom instruction and their professional growth. Skilled evaluators are essential to a valid and reliable educator evaluation process. However, North Carolina has much work to do to ensure educator evaluation is fair, reliable, and accurate. Little attention has been given to ensure that evaluators are trained to make judgments regarding a teacher's performance. There are several ways of framing this problem, each of which may dictate a different approach to a solution.

One way of framing the problem is through a policy lens. North Carolina policy regarding teacher evaluation provides little direction for Local Education Agencies (LEAs) to ensure their educators are qualified to evaluate. The only requirement for North Carolina evaluators is a North Carolina administrative license acquired by

completing a degree program in school administration. In order to be authorized by the state to offer an educator licensure program, an Institute of Higher Education (IHE) must submit a plan for assessing all candidates' competence in six standards, including "Teacher and Staff Evaluation." The variation in assessment requirements among all of the state's principal preparation programs, however, does not ensure uniform competence in evaluation skills for all entry level principals. Likewise, in-service principals and assistant principals are evaluated on eight standards, including teacher and staff evaluation. However, there is no requirement that practitioners submit samples of their evaluations as artifacts, leaving the process of assessing principal expertise to the supervisors' discretion. Currently, there is also no requirement for evaluators to demonstrate competence in using the North Carolina Educator Evaluation System (NCEES) to evaluate and provide feedback to licensed staff members, although this is one of the primary functions of the site-based administrator. Policy TCP-C-004: Policy establishing the Teacher Performance Appraisal Process Component 1: Training defines that "Before participating in the evaluation process, all teachers, educators and peer evaluators must complete training on the evaluation process." The consistency, quality, and fidelity of these trainings are left to the LEAs to initiate and complete. This neglect of pre- and in-service attention to the development of expertise in such a high-stakes task needs to be addressed through the policy process, as well as through state-provided support structures.

One such source of support for the adequate preparation and in-service development of educator evaluators might be expected from state-level policy leaders. A shift in policy that all administrative candidates become certified evaluators by

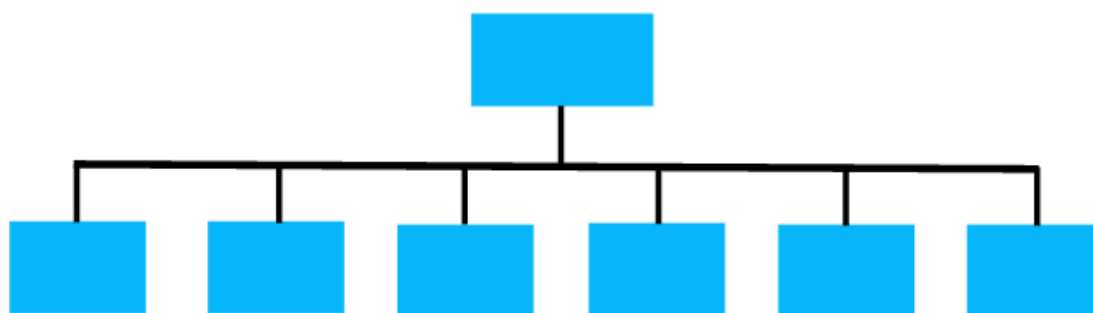
successfully completing a rater reliability evaluation program might improve both the accuracy of evaluation as well as the quality of feedback evaluators provide to teachers. Primary considerations for reforming the evaluation process would include policy changes regarding educator licensure, a partnership with institutions of higher education (IHEs), and an emphasis on supporting program implementation with fidelity through Observation Calibration Training (OCT).

Another lens for framing the problem of improving teacher evaluation is the human resource lens which focuses on what organizations and people do to benefit the other (Bolman & Deal, 2008). Schools are human organizations at the core, and school administrators are human resource managers, among their multiple roles. In a school setting, the capacity of principals to create an environment where teachers thrive requires human resource expertise. The human resource frame for teacher evaluation recognizes principals' roles in building teacher capacity. A principal's accurate evaluation of the individual teacher provides a platform for feedback to support growth and development. Principal human relations can inspire, motivate followers, and have the greatest impact on the behaviors and teacher performance, including their instructional abilities that result in student achievement.

Yet another problem framing lens is the North Carolina organizational lens, which provides insight into the support provided by the North Carolina Department of Public Instruction (NCDPI). Currently only one position is allotted to support principals and assistant principals in North Carolina. The position is North Carolina Educator Evaluation Consultant. This position is housed in the Educator Effectiveness division of DPI. Without additional personnel to develop and support the capacity of educator

evaluation skills, that responsibility falls to the local level to implement. Given this framework for capacity building, it makes sense for the state evaluation consultant to work collaboratively with the LEAs to provide principal development and support. This type of partnering for innovative 21st century professional development has the potential to reach all principals and assistant principals in the state.

The organizational lens for teacher evaluation within the actual school context can be viewed as a “flattened” hierarchy. Flat organizational structures do not have multiple levels of management between staff and the leader (Boundless, 2015). Flat structures ensure a level of responsibility from all employees and promote a diverse and creative platform for creative and empowering decision making. The direct relationship between teachers and principals as provided in a flat hierarchy provides a direct line of communication and feedback that is confidential (Boundless, 2015; see Figure 1).



Source: Boundless (2015), “Flat hierarchies.”

Figure 1. Flat organization chart.

A typical pyramid organization places all decision making and power with the leader. However, a school leader cannot effectively lead as a micromanager. The principalship is a complex position that must entrust teachers with shared decision making. Flat structures encourage teachers to make decisions to support growth for

themselves, success for their school, and ultimately student achievement. Fostering a collaborative culture in schools frequently complicates the role of the principal as an evaluator—a critical factor in addressing the problem of improving teacher evaluation through an organizational lens.

Research Evidence of the Problem

In order to improve the accuracy of teacher evaluations, many states, including North Carolina, have rolled out a new evaluation system for teachers and other educators in which value-added measures are used to compare teacher evaluation ratings with their students' measured and expected growth. Even with new evaluation systems in place, however, continued reform and development of such evaluation systems are needed, since in some states principals are rating 99% of their teachers as effective or better (Reform Support Network, 2013). Such inflated evaluation does not correlate with student performance data. In North Carolina, analysis of data on student achievement has revealed discrepancies in teacher evaluation ratings versus their corresponding value-added scores as determined by the Education Value-Added Assessment System (EVAAS) used. Thomas Tomberlin (2014), Director of District Human Resources Support at the NCDPI identified the correlation between each of the North Carolina Professional Teaching Standards (I-V) and the index, Standard 6. He found little correlation between teachers' Standard 6 ratings and teachers' ratings on the other standards 1–5. During the 2011–2012 school years, correlation was between .173 and .205. Correlation during the 2012–2013 school years was between 0.167 and 0.198. Interestingly, however, Tomberlin's (2014) report identified high correlation among and between each of the other five teacher evaluation standards (all around 0.70). The strong

correlation between standards 1–5 indicates that when educators evaluate teachers, they rate primarily the same on all five standards instead of considering each standard separately.

To examine possible causes of these discrepancies, I administered a survey to participants in NCDPI-hosted special training sessions called “Principal READY” meetings in each of the eight districts across the state during the spring of 2014. Participants identified reasons teacher evaluation ratings do not correlate with value-added ratings. Reasons included:

- Principals do not have a clear understanding of the North Carolina Standard Course of Study.
- Principals want to avoid conflict.
- Principals do not know how to have crucial conversations with ineffective teachers.
- Principals do not know how to coach teachers effectively.
- Principals have close relationships and personal ties to teachers and their families due to living in small communities.
- Principals are concerned that if they document and dismiss an ineffective teacher that they will not be allowed to fill the position due to budget cuts to education.
- Principals are concerned that if they terminate an ineffective teacher, the replacement might be even more ineffective.

Focusing on the possible workplace issues and repercussions for evaluating teachers negatively emerged as a set of possible explanations for the lack of congruence in ratings

of North Carolina principals. Additional explanations were identified by Robinson (2011) who described how judgments of educator quality are almost always independent of student learning. Her review of the literature enumerated three ways that educators are evaluated formally and informally. First, educators are frequently judged as effective if the school is well managed, safe, and clean, with sufficient academic supplies. Second, educators who are affable and social are judged to be effective. And third, educators who create positive relationships with politicians, parents, teachers, and students are assumed to be successful at their jobs, especially if they can appease the highly political school boards under whom they work.

To explore additional possible causes of rater discrepancies, I interviewed two district human resource leaders. These individuals suggested that principals were distracted by the new online evaluation tool platform (True North Logic) implemented during the 2013–14 school year. Both human resource directors identified the process of learning a new tool as being a distraction from the standards and from the implementation of the entire evaluation system. In fact, the time and energy that went into learning a new system resulted in principals spending little to no time on the feedback and the coaching process to support teacher growth. Furthermore, one of the interviewee's interpretations of the NCEES process was inaccurate. The interviewee said that a teacher can be doing something one year that is considered "proficient," but as expectations for that teacher change, the teacher could be viewed as "developing" the following year. This explanation is not aligned to the intended use of the instrument. Because human resource directors are responsible for supporting educators with NCEES, misconceptions such as this could account for one reason why many educators do not evaluate effectively.

Teachers should be rated against the standard, not against their previous performance. This misconception alone could account for all principals and assistant principals in a given LEA using the North Carolina teacher evaluation rubric incorrectly and ineffectively. Rater agreement is significant in ensuring that teachers receive similar accurate ratings on their performance.

Evaluation Training

Evaluation is not intuitive, but rather a purposeful act. Research suggests that evaluation is more effective when evaluators are trained (Darling-Hammond et al., 2011). Effective trainings should include resources that support the evaluation process (McGuinn, 2012). Coordinated, ongoing, sustained, professional development is the key to successful implementation of any evaluation system. According to McClellan, Atkinson, and Danielson (2012), educators need training for evaluation that addresses observer bias, provides opportunities to analyze and use the evaluation tool, and uses video of real classrooms to help educators gain a greater understanding of calibration (Guskey, 2002). The more practice educators have, the more likely they are to evaluate more accurately. Evaluation trainings vary greatly between states. Different formats for training delivery include train-the-trainer models, online modules that include directed and self-paced webinars, videos, and traditional face-to-face presentations.

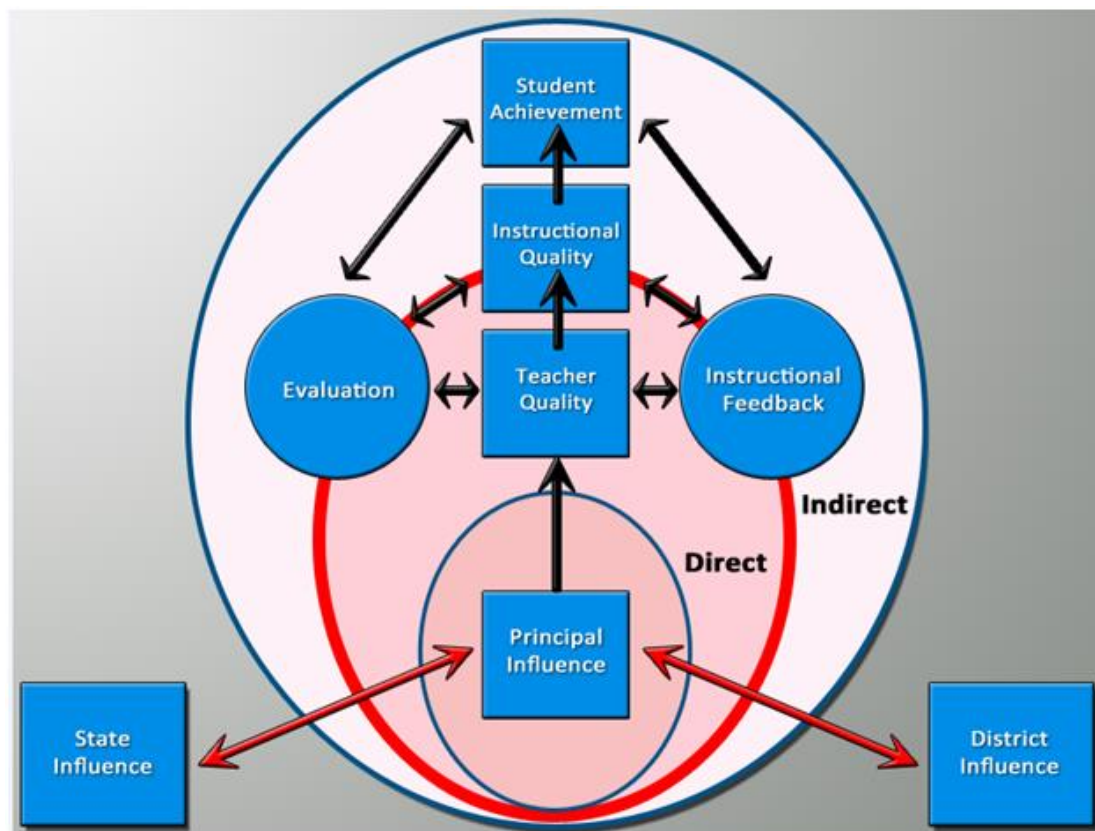
Since evaluation training is a relatively new concept, little data are available regarding the efficacy of the delivery format and content (Graham, 2011). Currently, three states (Illinois, Iowa, and Ohio) have instituted educator evaluation certification programs. These three programs each require teacher evaluators to engage in modules or trainings designed to improve the quality of feedback and inter-rater agreement. A

comprehensive study by Yoon, Duncan, Lee, Scarloss, and Shapley (2007) indicated that not only the duration of the professional development plays an important role in the success of the initiative, but that follow up and ongoing, job-embedded opportunities for discussion, feedback, and continued emphasis are all an integral part of a successful implementation (Darling-Hammond et al., 2011). Creating experiences for administrators where they can learn from reflecting on their experiences is an important part of the learning process. One way in which district leaders can ensure educators have the opportunity to reflect on their practice, specifically in terms of professional development implementation, is to provide follow-up sessions either as part of their regularly scheduled professional learning communities (PLCs) or in regularly scheduled follow-up sessions designed around reflection, sharing, and feedback. Providing such quality, ongoing professional development can not only change organizational patterns and norms but can lead to improved student outcomes (Guskey, 2000).

Conceptual Framework

The conceptual framework that best clarifies the complexity of teacher evaluation is a modification of “The Ripple Effect” framework used in a 2012 report published by The American Institutes for Research. The Conceptual Framework as shown in Figure 2 identifies the partnership between NCDPI and LEAs. At the onset of this project, this framework was used to suggest the collaboration between state leaders and district leaders could have a significant influence over a principal’s ability to evaluate and coach teacher growth resulting in student achievement (Clifford, Behrstock-Sherratt, & Fetters, 2012). By using this framework to explore aspects of teacher evaluation on which to focus improvement efforts, it is also clear that any change made to one component is

bound to affect other components, as well. Whether changes result in improvements was the question at the heart of this project. Furthermore, delivering development and support through a state and district collaboration was viewed as a means to address both the broad goals of the state agency and the local improvement goals.



Source: Simmons, K., & Mullins, H. (2013). Adaption of "The Ripple Effect." Retrieved from <http://edlstudio.wikispaces.com/file/view/ripple.png/527174698/ripple.png>

Figure 2. Conceptual framework.

Desired State

In light of the discrepancies between teacher evaluation data and student achievement data, the NCDPI sees a need to ensure that principals and assistant principals understand their role as evaluators and possess the skills needed to evaluate

teachers accurately. Because discrepancies were identified across the state, the need for a statewide professional development plan and calibration system became apparent.

Calibration of the evaluation rubric used to rate teacher performance based on the five observable standards from the North Carolina Professional Teaching Standards is a process in which every administrator responsible for evaluating teachers need to engage, in order to ensure greater accuracy among all evaluators. Performance review calibration measured against a standard, improves the reliability in a rating system for evaluation and promotes honesty and fairness in the ratings system for evaluation of teachers.

Calibration as professional development is a process in which multiple observers or evaluators collaborate and discuss performance ratings based on objective evidences (Performance Review Calibration, n.d.). The calibration process provides a better understanding of the professional teaching standards and instructional practices.

Ensuring a high level of understanding of standards and practices for all evaluators is the desired state of the North Carolina Educator Evaluation System.

CHAPTER TWO: SETTING AND CONTEXT OF THE IMPROVEMENT PROJECT

Historical and Current Initiatives

Since the 1980s, North Carolina has used a statewide teacher evaluation tool. In 1997, the North Carolina State Board of Education charged the North Carolina Professional Teaching Standards Commission with the alignment of the Core Standards for the Teaching Profession with the newly adopted State Board mission that every public school student will graduate from high school globally competitive for work and postsecondary education and prepared for life in the 21st century. To this end, Commission members considered what teachers need to know and be able to do in 21st century schools (Yoon et al., 2007). The commission's work resulted in a set of criteria for evaluating the state's teachers.

In February 2007, the NCDPI partnered with Mid-continent Research for Education and Learning (McREL), a private nonprofit organization, to establish a new instrument for statewide evaluation known as the NCEES (North Carolina Educator Evaluation System). The North Carolina State Board of Education adopted five aligned standards in June 2007 (Ho & Kane, 2013). In October 2008, the North Carolina State Board of Education approved the policy adopting the rubric for evaluating North Carolina teachers and the teacher evaluation process. The state rolled out NCEES to districts across the state in phases. Thirteen of 115 school districts implemented the new rubric beginning in the 2008–2009 academic year (McGuinn, 2012). Half of the remaining districts implemented the new system beginning in 2009–2010, and the final group of LEAs adopted the system in the 2010–2011 school year. The evaluation includes five adopted standards, 25 elements, and four distinct groups of descriptors for the four

ratings of each element. Teachers can be rated as developing, proficient, accomplished, or distinguished on the 25 different elements that fall under one of the five standards.

In February 2012, the North Carolina State Board of Education adopted a sixth standard: Teachers Contribute to the Academic Success of Students, as part of teacher and educator evaluations during the 2011–12 school years (Simmons & Mullins, 2013). North Carolina uses student growth data as measured by the EVAAS from the SAS Institute to determine the effectiveness of teachers, schools, and districts with regards to student achievement and provides an analysis of student performance on standardized assessments and related expected teacher performance. EVAAS provides teachers and educators with multiple resources and reports dissecting the performance of student data. Because high stakes decisions are linked to teacher evaluation, additional concern has been raised about the lack of correlation between teacher ratings based on educator evaluation and student learning outcomes (Reform Support Network, 2013).

A multi-factorial correlation study between teacher performance evaluation data and the EVAAS student achievement data for 2010–2011 did not find a correlation between performance evaluation data and EVAAS data. The dataset included 11,430 North Carolina teachers in 35 LEAs having both EVAAS scores and performance evaluation ratings assigned in the 2010–11 school year. Although 46,000 teachers had evaluation data for 2010–11, only around 11,000 of those also gave an End of Grade (EOG) or End of Course (EOC) assessment. These 11,000 received an EVAAS data score (NGA Center for Best Practices, 2011).

Research found that there was a small distribution of evaluation ratings in the study. Out of 11,000 teachers, the 100 teachers with the best student achievement data

received the same ratings on their evaluation as the 100 teachers with the worst achievement data. The study did not find a correlation between performance evaluation data and EVAAS data. Rowan-Salisbury had the strongest correlation of the 35 LEAs in the study (Batton, Britt, DeNeal, & Hales, 2012). High Stakes decisions like teacher status and compensation pay have highlighted the validity of the teacher evaluation.

Recent Initiatives and the Call for Improvement in Teacher Evaluation

Currently, the North Carolina Public School system serves approximately 1,520,305 students. There are 2,434 schools among 115 LEAs and 148 Charter Schools. Education funding is roughly 37% of the NC general fund. EOG and EOC test scores, along with school accountability growth, is calculated using a value-added growth tool (EVAAS). Each school is designated as having exceeded growth, met growth, or not met growth. The results for school accountability growth are shown in Table 1.

Table 1

2014–15 School Accountability Growth

Growth Category	Number	Percent
Exceeded Expected Growth	689	27.6%
Met Expected Growth	1,116	44.7%
Did Not Meet Growth	691	27.7%

Note. Adapted from 2014–15 Performance and Growth of North Carolina Public Schools Executive Summary (NCDPI, 2015a, p. 1)

We have 98,544 teachers in North Carolina, including charter school teachers who serve our students while earning an average annual salary of \$47,783, compared to the national annual average of \$57,379 (NCDPI, 2015b). Eyewitness News at 11 reported in March 2015 that North Carolina teachers have moved from 47th in the nation

for teacher pay to 42nd in the nation. The Governor's budget for increased teacher salaries provided salary increases for only one-third of the state's teachers. And, North Carolina is ranked among the lowest in the nation with per pupil spending that decreased from \$8,632 in 2014 to \$8,620 in 2015 (Waliga, 2015).

Even in bad economic times for NC schools, teachers and students are working harder than ever to keep up with the demands of a 21st century learning environment. In 2014, the Southern Regional Education Board (SREB) published a state report titled "North Carolina: Taking Stock and Pushing Forward, 2014." The report identified notable student outcomes in NC:

- North Carolina's state-funded pre-K program met all 10 of the nationally recognized standards of quality, one of four programs in the nation to do so.
- North Carolina's fourth graders *outperformed* the nation in reading and math achievement on NAEP at the Basic and Proficient levels and *outpaced* the nation in reading at the Basic level. Eighth graders outperformed the nation in math achievement on NAEP at the Basic and Proficient levels and outpaced the nation in reading at the Basic level.
- North Carolina's high school graduation rate outpaced the nation in growth. This increase in graduation rate extended to black, Hispanic and white high school seniors.
- North Carolina's six-year graduation rate for first-time, full-time freshmen who entered public, four-year colleges and universities topped the national and regional rates. (SREB, 2014, pp. 1–2)

In addition to student achievement, North Carolina has committed to implementing the NCEES statewide. The NCEES contains six evaluation standards for teachers. The evaluation standards are:

1. Teachers demonstrate leadership
2. Teachers establish a respectful environment for a diverse population of students
3. Teachers know the content they teach
4. Teachers facilitate learning for their students
5. Teachers reflect on their practice
6. Teachers contribute to the academic success of students

Standard 6 is based solely on quantitative measures of student growth. Standards one through five are based on educator observations of teachers, primarily observations that take place in teachers' classrooms. As part of Race to the Top, North Carolina agreed to provide training to educators and teachers on the evaluation system. The NCDPI provided professional development tools and opportunities geared toward each element of the evaluation system. Trainers discovered disconnect between training about the evaluation system in the abstract and actually allowing observers to practice conducting observations during training. As a result, the NCDPI recognized the need for an online platform for observer evaluation training aligned to the six professional teacher evaluation standards as seen above.

NCDPI contracted with BloomBoard, Inc. to provide a platform Observation Engine for professional development and calibration for teacher evaluators. The OCT provides a suite of training activities for North Carolina administrators to improve their

accuracy and reliability when observing and evaluating teachers. The setting for this project included 20 LEAs in North Carolina that are participating voluntarily in a professional development pilot, Observation Calibration Training (OCT). Approximately 360 participants including elementary, middle, and high school principals; assistant principals; and central office administrators volunteered to engage in virtual professional development.

Observation Engine is the online platform that houses the training activities, Scoring Studies and Lessons, to be completed during the training. Scoring Studies help build consensus and inter-rater reliability among a group of evaluators. A Scoring Study assigns a video to observers who must watch and rate the video using the NCEES rubric. A Scoring Study report provides helpful information about observer agreement with both target and modal scores, as well as the general distribution of scores across a group of observers. For this project Scoring Studies were used at the beginning and at the end of the pilot to measure improvement as a result of the activities during the pilot, much like pretests and posttests. Lessons provide targeted, self-paced online learning activities for evaluators. Designed for professional development activities associated with the NCEES rubric, lessons provide immediate on-screen feedback for observers that appears as soon as they have submitted their scores. The OCT aims to improve observation skills, increase rater agreement, and to provide a common experience for LEAs to host collaborative conversations to improve instructional leadership skills.

BloomBoard, Inc. used a team of six subject matter experts with backgrounds in education as instructors, educators, and educational methodology trainers to select six videos as examples of teaching and student instruction. The six videos were chosen from

a pool of 24 videos such that two videos were selected for each of three relevant instruction levels (grades 3–4, 6–7, and 8–9). For each instruction level, one video was selected that illustrated English language instruction and another video was selected that displayed mathematics instruction.

Each subject matter expert, over the course of one week, rated each video on the four standards and the 17 associated rubric descriptors for the standards. The subject matter expert team was then reconvened and all ratings were reviewed for consensus. All ratings that did not achieve initial consensus among the subject matter experts were carefully reviewed, discussed, and a final consensus rating was given to the descriptor until all descriptor ratings for each video achieved consensus.

In addition, a team of three NCEES consultants reviewed each of the 34 video lessons and two full length classroom videos and scored each teacher using the professional teacher rubric. The scores were then cross-referenced with the scores provided by the master scorers from BloomBoard, Inc. Final scores were determined collaboratively with justifications for each score. With this pilot, North Carolina educators can access observer calibration events to practice teacher observation, increase rater agreement, and promote continuous improvements to the NCEES.

On November 19, participants were provided with an introduction to the project via a live webinar. They were also given access to written instructional materials and a short demonstration video. Participants were instructed to first complete Scoring Study 1 (which served as a “pretest”) and to then complete the 19 observable lessons over the course of approximately five months at their own pace. Scoring Study 2 was then

administered as a “posttest.” This project will focus specific interventions occurring post Scoring Study 1.

CHAPTER THREE: INTERVENTION DESIGN

The goal of this project was to develop a sustainable process to improve rater performance on the North Carolina Professional Teacher Standards 1 through 5. Information and data gathered from the project will also inform phase 2 of the Observation Calibration Training (OCT) statewide training plan with the overarching aim of improving participants' teacher rating competence. Eventually the online training platform will become a component of the statewide educator evaluation system training program and may later be used by districts or the Department of Public Instruction to review and certify observers/evaluators. We expected each intervention to provide data informing each subsequent adjustment and addition to the pilot project. This chapter presents an overview of the improvement science methods and strategies developed and implemented to achieve the goal.

Intervention Design Team

The design team members from NCDPI were assigned by the Chief Academic Officer and the Director of Educator Effectiveness for NCDPI. As the North Carolina consultant for the educator evaluation system for the past three years, I brought educator and leadership experience to the position. I host regional Principals Council meetings across North Carolina and my division, Educator Effectiveness, hosts Principal Ready meetings in the fall and spring. The OCT pilot is funded through Race to the Top and the Project Coordinator, along with two professional development leads with educator experience, rounded out the design team from NCDPI.

NCDPI contracted with BloomBoard, a California-based educational development company founded in 2010. BloomBoard enables personalized professional development

for K–12 educators by creating individualized professional growth opportunities, resources, and services specific to the needs of individual educators. BloomBoard provided

- Observer Calibration Events & Reporting: practice with rating and help with increasing rater agreement with target scores
- Supplemental Learning Exercises: targeted, self-paced online exercises focused on specific NCEES elements (BloomBoard + Empirical Education, n.d., p. 1)

BloomBoard partnered with Empirical Education, a Silicon Valley-based research company that provides tools and services to help K-12 school systems make evidence-based decisions about the effectiveness of their programs, policies, and personnel. The company brings its expertise in research, data analysis, engineering, and project management to customers that include the U.S. Department of Education, educational publishers, foundations, leading research organizations, and state and local education agencies. (BloomBoard + Empirical Education, n.d., p. 2)

Bloomboard professional development resources include:

- Access to a marketplace of articles, videos, and video clips, as well as links to web-based resources
- Supplement and support the work educators are doing through Observation Engine’s scoring studies and calibration (BloomBoard + Empirical Education, n.d., p. 1)

BloomBoard and Empirical Education are partnering with NC to provide calibration and training for administrators across the state. BloomBoard and Empirical

Education's team members included two project coordinators for the OCT and contracted specialists from McRel International. McRel partnered with North Carolina to develop the NCEES (North Carolina Educator Evaluation System). Two of the McRel-contracted scorers were part of the team who wrote and conducted initial testing of NCEES.

Collectively, master scorers brought the current desires from administrators in the field along with a strong historical memory of the NCEES. Each team member has an educational background with a focus on educator effectiveness and educator evaluation.

The OCT interventions were determined by BloomBoard's OCT project lead, Empirical Education's project lead and myself, NCDPI NCEES Consultant. I wrote the charge to the team and shared it at the onset of the OCT development. The charge was:

- The vendor will assist the NCDPI in the development of an online observation platform that meets certain desired criteria, including:
- A platform to support videos and a master scoring rubrics that NCDPI can calibrate and use to train observers in the evaluation system.
- This tool must be a clear reliable product which can be implemented and utilized uniformly statewide.
- In order for the NCDPI or districts to conduct deeper analyses of individual observers, or trends in observer tendencies, the observation platform must have reporting capacity.

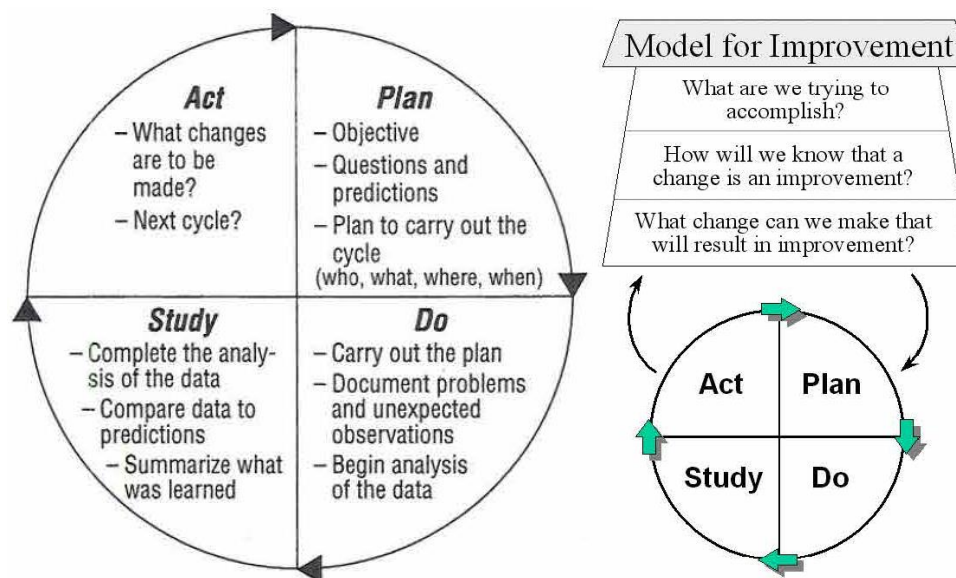
Intervention Process

Overview of the Improvement Model

The improvement project involved implementing an online training platform and monitoring the progress of 360 participants in a pilot of the state's web-based

Observation Calibration Training (OCT) during a seven-month period. Formative assessments included monthly participation and performance data reports, to ensure that the OCT pilot was progressing and to identify elements that participants were scoring correctly and incorrectly. Based on this data, I developed and facilitated webinar intervention cycles to address rating performance issues. The project intervention cycles and Plan, Do, Study, Act (PDSA) Cycles are analogous, providing a basic framework for developing, testing, and implementing changes that will lead to improvement. (Langley, et. al., 2009, Moen, 2015). Figure 3.1 identifies the PDSA Cycle as a continuous process to support and inform what is trying to be accomplished, how we know that a change is an improvement and changes we can make that may result in improvement. This model for Improvement removes the quest to find a solution; instead it provides permission to try new ideas. If an idea works, then you can expand and extend the change. If it doesn't work, you can adjust or abandon without having wasted extensive resources without a solution.

The intervention cycles included systematic steps to inform the continual improvement of teacher evaluators rating teacher performance accurately. The PDSA format ensured that intervention implementation was based on formative data and informed the progress and analysis of an intervention cycle. The approach supported a full range of improvement efforts from the very informal to the most complex.



Source: Moen (2015), p. 8.

Figure 3. PDSA Cycle and Model for Improvement, 1991, 1994.

Data Sources and Uses

The OCT training included two scoring studies, lasting approximately 50 minutes. These served as pre and post data to measure rater competency prior to and after completing the entire seven months of training lessons. Scoring Study results were provided to Local Education Agency (LEA) pilot leaders and individual participants following the completion of the Scoring Study. Scoring Study results informed the LEAs, conducting facilitated activities, which elements their participants needed additional understanding or support. Individuals were encouraged to use the information from their Scoring Study report to identify the elements they scored incorrectly, then complete lessons for those identified elements first in their training. Individual participants had flexibility to complete element lessons in any order they chose. Some LEA group participants were given the same flexibility while other chose to create a

schedule of completion. This ensured a shared context for deep dive discussions of the element specific video clips.

Intervention Procedures

Lessons within the OCT included two full video lessons lasting an entire class period and 17 mini lessons that were element specific. The full videos were of complete class lessons. All 17 observable elements were scored at the conclusion of each full video lesson. The element specific lessons are the 17 observable elements identified within the NC Professional Teacher Standards. Each of the 17 elements provided two 5- to 7-minute video clips which pilot participants scored using the state-adopted scoring rubric. Both full videos and the 34 video clips could be watched multiple times before scoring. This practice is different from an authentic classroom setting where an observer only has one chance to observe evidence for a score. In some instances, numerous viewings for practice were necessary for confirmation of questionable evidence. Also, with online training, viewing may be interrupted by technical or other distractions. Scoring results and justifications for scores were provided to participants immediately following the input of lesson scores for full video lessons and element specific lessons.

The complete menu of data sources includes:

- Scoring Study Reports:

NC Scoring Study 1 Aggregated Report

Distribution of Modal and Target Scores by Video

Distribution of Ratings by Domain

Observer Agreement with Target Scores by Domain

NC Scoring Study 2 Aggregated Report

Distribution of Modal and Target Scores by Video

Distribution of Ratings by Domain

Observer Agreement with Target Scores by Domain

- Periodic Participation Reports
- Periodic Performance Reports
- Informal feedback collected during webinars
- End of year Participant Survey
- Focus Group Feedback
- Concordia Schools Case Study Report

These data sources, drawn from one or more of the intervention phases, informed each modification in the implementation process.

Data Analysis

In monitoring and modifying the state's pilot Observation Calibration Training program, I used improvement science methods, including quantitative methods that enabled me to analyze the strategies that were useful in improving the participants' evaluation knowledge and skills.

To validate the Observation Calibration Training, I analyzed data collected from two Scoring Studies serving as pretest and posttest. The scoring studies each consisted of a 45-minute video that had been scored by a team of six expert researchers using the 17 observable elements from the North Carolina Educator Evaluation System (NCEES) rubric. A paired sample *t*-test was used to compare the means of a normally distributed interval dependent variable for two independent groups. The participants in each group

were the same participants, making them related. This is a repeated measures design for a quantitative measure of participant rater accuracy.

Quantitative data analysis on the pretest and posttest occurred at the beginning and end of the pilot. Key metrics for the scoring study reports included Percent Target Agreement, the percent of scores that agree exactly with the target score and Percent Target Discrepant, the percent of scores that disagree with the target scores by two or more performance levels. More focus was given to the end of the year to identify rater improvement by participants scoring the videos of teachers' classes.

Periodic assessment of change and analysis of progress toward improvement included monthly performance data reports ensuring that the pilot was progressing for participants and to identify the elements from the Professional Teacher Standards that participants were scoring incorrectly. Based on these data, optional webinars were developed for participants to attend. The webinars provided support for participants through clarification and additional insight into evidences of the individual elements. Best practice to get the greatest benefit from the OCT was also shared. After each intervention cycle, the design team reviewed the data informing the OCT and made adjustments in the next cycle, as described in Chapter 4.

In addition to quantitative data, qualitative data including Focus Group Data, Participant Survey Data, and the Concordia Schools' Data were used to inform improvement goals moving forward from the pilot to the first year of implementation in the 2015–16 school year. Participant insight into the online tool, content, and overall support provided a valuable user perception that clearly identified strengths and weaknesses for the overall project.

Implementation Plan Including Projected SMART Goals That Spanned the Project

Table 2

Implementation Plan

Project Purpose/Goal	To develop a sustainable process to improve rater performance on the NC Professional Teacher Standards 1 through 5.
NCDPI Goals	Increase observation skills resulting in stronger scoring calibration.
Objectives	<ul style="list-style-type: none"> • Gather pre/post calibration data using tools (scoring study/test) in Observation Engine to demonstrate increased calibration. • Develop observer capacity through strategized professional learning interventions with webinars and resources. • Determine appropriate interventions to increase rater agreement with the North Carolina Educator Evaluation System.
Duration	November 2014–June 2015
Project Smart Goals	<ul style="list-style-type: none"> • North Carolina teacher evaluators will participate in the 2014-2015 OCT Pilot so that rater agreement will increase by June 2015 as measured by a comparison of Scoring Study 1 and Scoring Study 2. • North Carolina teacher evaluators will participate in the 2014-2015 OCT pilot so that rater discrepancy will decrease by June 2015 as measured by a comparison of Scoring Study 1 and Scoring Study 2.

List of Supporting Artifacts

The following artifacts used during the implementation of the project can be found in the Appendix:

- OCT Participant Guide
- Observation/Evaluation Rubric
- Observation/Evaluation Rubric—Fillable

- OCT Kick-off Webinar Presentation
- Questions for Post Observation Conferences and Summative Evaluation
- Evidences for Professional Teacher Standards 1–5

CHAPTER FOUR: IMPLEMENTATION AND RESULTS

The ultimate goal of this project was to develop a sustainable process to improve rater performance on the North Carolina Professional Teacher Standards 1 through 5. The immediate goal was to improve the delivery and outcomes of participants in a voluntary online training program. The improvement process included three intervention cycles designed to increase participant observation and rating skills. Within the online Observation Calibration Training (OCT), Scoring Study 1 and Scoring Study 2 served as pre- and post-data in Observation Engine to demonstrate increased calibration. Strategized professional learning interventions with webinars and resources were used to develop participant capacity and increase rater agreement with the North Carolina Educator Evaluation System (NCEES). In this chapter, I will share each intervention phase of the Observation Calibration Training implementation and explain how data analysis and outcomes affected actions that led to each subsequent improvement cycle. The intervention framework illustrated in Figure 4 provides an outline of the improvement cycles in this study.

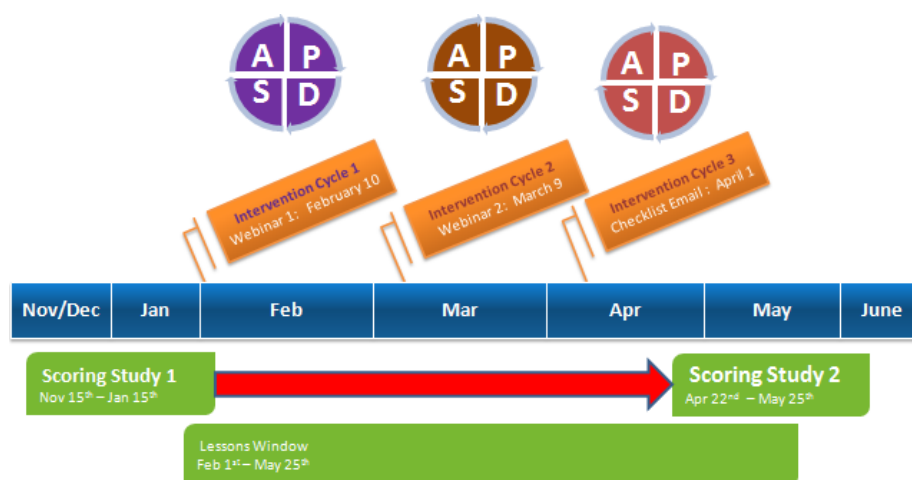


Figure 4. Intervention framework.

Intervention Cycles

Pre-Cycle: Scoring Study 1

Identification of areas of disagreement among raters. Following their participation in the introductory webinars in November and December, OCT participants were supposed to complete the first Scoring Study, “Scoring Study 1” which included watching a 50-minute video of a full class period and score the teacher according to the 17 observable elements from within the NCPTS rubric for observation and evaluation. Of the 360 registered participants, 83 had completed Scoring Study 1 by the original deadline of January 1, 2015. The deadline was therefore extended to January 15th, 2015 to allow more participation. A reminder email was sent to all registered participants. A few districts requested further extension. By February 5th, 2015, 128 participants had completed Scoring Study 1. Empirical Education provided the North Carolina Department of Public Instruction with a NC Scoring Study 1 Aggregated Report. Video V209 was used for Scoring Study 1. The report included a distribution of modal and target scores by video. Based on the data, I identified five elements that were most frequently scored incorrectly. The most frequently incorrectly scored elements were:

1. Standard 2c Teachers treat students as individuals.
2. Standard 2d Teachers adapt their teaching for the benefit of students with special needs.
3. Standard 3a Teachers align their instruction with the North Carolina Standard Course of Study (NCSCS).
4. Standard 3d Teachers make instruction relevant to students.

5. Standard 4a Teachers know the ways in which learning takes place, and they know the appropriate levels of intellectual, physical, social, and emotional development of their students.

Empirical Education also provided a “Distribution of Ratings by Domain” Report for video V209. These data determine the agreement between pilot participants as they scored the video. This information helped me identify the elements with closely related rating scores. For example, Standard 3c: Teachers recognize the interconnectedness of content areas/disciplines, was rated “developing,” the correct target score, by 47 of the 128 participants (37%) who completed Scoring Study 1. However, 47 participants also rated the same teacher in the video as “proficient.” This indicated to me that there was strong disagreement between the “developing” and “proficient” ratings for this video.

Elements that observers strongly disagreed on were:

1. Standard 3c Teachers recognize the interconnectedness of content areas/disciplines.
2. Standard 3d Teachers make instruction relevant to students.
3. Standard 4b Teachers plan instruction appropriate for their students.
4. Standard 4g Teachers communicate effectively.

The interesting factor in this data set was that the disagreement for all of the identified elements was between “developing” and “proficient.” My experience working with administrators across the state has identified more of a discrepancy between ratings “proficient” and “accomplished.” I suspected the nonexistent relationship between the participant and the teacher in the video being rated influenced the deflated category of ratings. In fact, research suggests that evaluators are more lenient when they know they

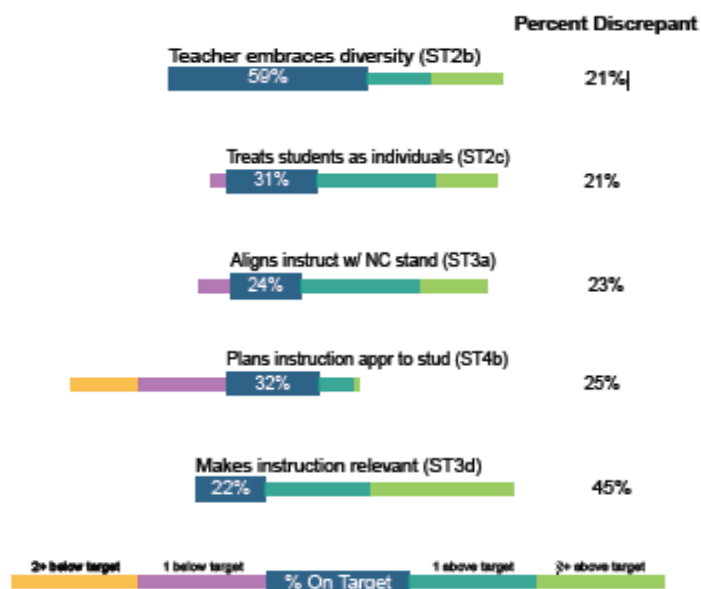
will have to provide justification for their ratings in a face-to-face meeting with the person being evaluated (Levy & Williams, 2004). Additionally, people develop relationships when they work together, which tends to bias observers (Antonioni & Park, 2001). I concur with scholars who have found that ratings are inflated when observers have an established relationship with the teacher.

Identification of participation related to performance. Periodic assessment of change and analysis of progress towards improvement included participation and performance reports. Data were used to ensure that participants were progressing and to identify the elements, from the Professional Teacher Standards, that participants were scoring incorrectly. These data indicated a need for additional support for participants in the OCT. I (or my team and I) used that data as a basis for developing optional webinars (intervention cycles) for participants to attend.

Participant Agreement with Target Scores by Domain data informed me of the elements that challenged participants during their scoring exercise (see Figure 5). The bars indicate the percent rated on target and the percentage of scores that disagreed with the target score. Exact discrepancies indicated more than 1 point variance is also provided. I also used this discrepancy data graph to determine if raters were scoring above the target score or below the target score. I identified five standards with high discrepancy. Four of the five elements were discrepant above the target score and one was 25% discrepant below the target score.















Using the three data sources, “Distribution of Modal and Target Scores by Video,” “Distribution of Ratings by Domain,” and “Observer Agreement with Target Scores by Domain,” I identified the elements which were scored incorrectly, were least


agreed upon, and had the greatest discrepancy from the target score. This combined information in Figure 6 informed Intervention 1 of my improvement cycle. Based on Figure 6, Element 3d Teachers make instruction relevant to students, is recognized by all three data sources as challenging elements and in need of further study.




Adapted from Empirical Education (2015), pp. 8–9.

Figure 5. Participant agreement with target scores by domain.

Standard 1	Standard 2	Standard 3	Standard 4	Standard 5
a. Leads in the classroom	a. Provides an environment that is inviting, respectful, supportive, inclusive and flexible	a. Aligns instruction with the North Carolina Standard Course of Study  	a. Knows the ways in which learning takes place, and the appropriated levels of intellectual, physical, social, and emotional development of students 	a. Analyzes student learning
b. Leads in the school	b. Embraces diversity in the school community and in the world 	b. Knows the content appropriate to the teaching specialty	b. Plans instruction appropriate for students  	b. Links professional goals
c. Leads the teaching profession	c. Treats students as individuals  	c. Recognizes the interconnectedness of content areas/disciplines 	c. Uses a variety of instructional methods	c. Functions effectively in a complex, dynamic environment
d. Advocates for the school and students	d. Adapts teaching for the benefit of students with special needs 	d. Makes instruction relevant to students   	d. Integrates and utilizes technology in instruction	
e. Demonstrates high ethical standards	e. Works collaboratively with families and significant adults in the lives of their students		e. Helps students develop critical-thinking and problem-solving skills	
			f. Helps students work in teams and develop leadership qualities	
			g. Communicates effectively 	
			h. Uses a variety of methods to assess what each student has learned	

 Distribution of modal and target scores by video (elements most frequently scored incorrectly)

 Distribution of Ratings (elements strongly disagreed on by participants)


 Observer Agreement with Target Scores by Domain (elements with high discrepancy)

Figure 6. Scoring Study 1 identified elements for further study.

Intervention Cycle 1: Webinar

Intervention 1 webinar agenda on February 10th included:

- Review Scoring Study
- Prioritizing Lessons
- Tips on Improving Calibration
- Rubric Study
- Scripting and Aligning Evidence

Webinar 1 was planned as an overview of Scoring Study #1 results, including a calibration discussion of participants' scores alignment and non-alignment with the master scores. In response to Scoring Study 1, I recognized a need to include specific

support of Element 3d. I worked collaboratively with the project leads from Empirical Education and BloomBoard to design webinar interventions to improve the scoring capacity of participants with particular attention to Element 3d. The webinar was optional for the OCT pilot participants. Twenty-nine of the 360 registered participants registered for the webinar and 22 attended it. However, it is difficult to determine how many participants actually attended the webinar, since some LEAs and charters attended the webinars as a group and registered using one registration. The webinar was also recorded to allow for later viewing for those who couldn't attend, and no registration was required for those who viewed the webinar at a later date. Several LEA leads requested aggregated scoring reports for the participants in their district. Intervention 1 webinar included additional information about accessing the participants' individual results within the online pilot tool.

Strategies for increasing personal accuracy. Webinar leaders, including the two project leads and I provided strategies for increasing personal accuracy on calibration. Strategies included an Indicator Study of the specific elements participants scored incorrectly. The Indicator Study involved a dissection of the elements to identify evidences to indicate proficiency on the element. In addition to proficiency, the indicator study examines the rating categories and identifies possible indicators related to each rating.

Element 3d Teachers make instruction relevant to students was highlighted during Webinar 1 as shown in Figure 7. Figure 7 lists scores that transfer to the ratings as:

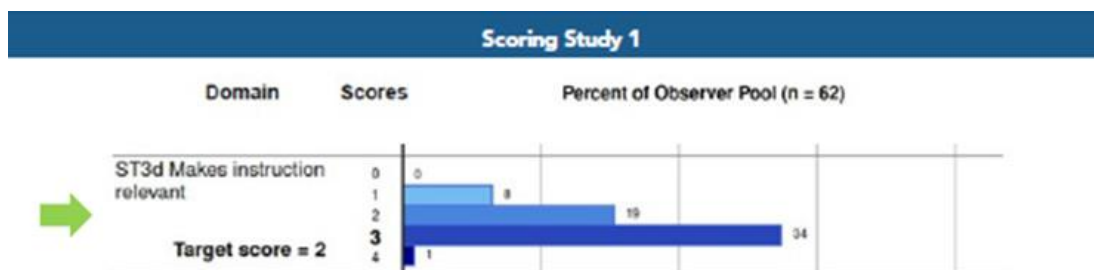
0 = Not Demonstrated

1 = Developing

2 = Proficient

3 = Accomplished

4 = Distinguished



Source: Empirical Education (2015), p. 11.

Figure 7. Analyzing disagreement: Sample scoring distribution for ST3d.

Figure 7 shows that 34 of the 62 participants who had completed this scoring thought the teacher should have been scored a 3 (Accomplished) rather than the target score of 2 (Proficient). The distribution of ratings for Standard 3d provided a relative example for the indicator study during the first webinar intervention.

Scripting quality evidence. Another strategy was Scripting Quality Evidence. This strategy recommended the observer script recognized evidence during an observation. In addition to the indicator study with Standard 3d, we also brainstormed possible quality evidence, from Standard 3d, that would warrant scripting. I specifically mentioned that a rating of distinguished for 4d, described as “deepens students’ understanding of 21st century skills and helps them make their own connection and develop new skills,” would require evidence of the students’ behavior. Standard 3d was integrated into the webinar as context for the Indicator Study and Scripting Quality Evidence.

Developing a personal plan. Webinar leaders also encouraged a personal plan for prioritizing the OCT lessons in order to increase calibration rates. Webinar participants were advised to review their personal scoring study results and compare them to the true scores in the scoring study report. They were encouraged to create a prioritized list of elements based on their least calibrated areas. Other considerations for prioritization included unclear proficiency criteria and indicators that were difficult to rate but still calibrated. Standard 3d, an element identified in the scoring study as incorrectly scored least agreed upon with the greatest discrepancy from the target score was reviewed and specific indicators were discussed for rating clarification.

Monthly monitoring of participation and performance. Proposed progress monitoring for Intervention 1 was to include monthly reports on participation and performance, including identification of standards not rated accurately by participants. In addition, a redistribution of the participant survey including perception data by participants was scheduled. The participant and performance data were analyzed. The participant survey scheduled for the end of the pilot was maintained to collect summative data for the OCT implementation the following year.

In addition to the Scoring Study 1 performance data, I analyzed progressing participant performance data including element-specific lesson performance. Elements scored incorrectly most often on lessons through February 19, 2015 are shown in Table 3.

Table 3

Elements Scored Incorrectly Most Often 2/19/15

Element II d	<p>Teachers adapt their teaching for the benefit of students with special needs.</p> <p>Teachers collaborate with the range of support specialists to help meet the special needs of all students. Through inclusion and other models of effective practice, teachers engage students to ensure that their needs are met.</p>
Element III a	<p>Teachers align their instruction with the <i>North Carolina Standard Course of Study</i>.</p> <p>In order to enhance the <i>North Carolina Standard Course of Study</i>, teachers investigate the content standards developed by professional organizations in their specialty area. They develop and apply strategies to make the curriculum rigorous and relevant for all students and provide a balanced curriculum that enhances literacy skills. Elementary teachers have explicit and thorough preparation in literacy instruction. Middle and high school teachers incorporate literacy instruction within the content area or discipline.</p>
Element III d	<p>Teachers make instruction relevant to students.</p> <p>Teachers incorporate 21st century life skills into their teaching deliberately, strategically, and broadly. These skills include leadership, ethics, accountability, adaptability, personal productivity, personal responsibility, people skills, self-direction, and social responsibility. Teachers help their students understand the relationship between the <i>North Carolina Standard Course of Study</i> and 21st century content, which includes global awareness; financial, economic, business and entrepreneurial literacy; civic literacy; and health awareness.</p>
Element IV a	<p>Teachers know the ways in which learning takes place, and they know the appropriate levels of intellectual, physical, social, and emotional development of their students.</p> <p>Teachers know how students think and learn. Teachers understand the influences that affect individual student learning (development, culture, language proficiency, etc.) and differentiate their instruction accordingly. Teachers keep abreast of evolving research about student learning. They adapt resources to address the strengths and weaknesses of their students.</p>

Addressing low participation in the online training. While much of my attention was on identifying elements that were not being scored correctly, it became overwhelmingly evident that participation in the pilot was low with only 25% of participants completing a task in the OCT in seven months. A reminder email was sent out to all 360 active participants on March 1, 2015. On March 4th, 2015, only 118 of the 360 participants had opened the email and 17 users clicked one of the links. Only 90 of 360 registered users had completed a task in the Observation Calibration Training by March 8th. I decided that increasing participation would be an additional goal of my next intervention.

Intervention Cycle 2: Webinar

Intervention 2 webinar agenda on March 8th include

- Setting Up Successful Structures: Best practices for facilitating PD sessions
(Heather Mullins, Chief Academic Officer, Concordia Schools Data)
- Trends
- Coaching Resources
- Next Steps

Progress monitoring clearly identified the low number of engaged participants as a data point to inform the content for the Intervention 2 Webinar. In addition to low participation, specific elements scored incorrectly most often needed to be examined to provide greater insight and clarity around the specific evidences to support the elements. I decided to choose an element from Standard 2: Teachers establish a respectful environment for a diverse population of students, Standard 3: Teachers know the content they teach, and Standard 4: Teachers facilitate learning for their students. I

did not choose Standard 1: Teachers demonstrate leadership, because there is only one observable element in Standard 1. I did not choose Standard 5: Teachers reflect on their own practice due to no observable elements within Standard 5. Standards 2, 3, and 4 focus on what is happening in the classroom, the content being taught, and how it is being taught. I chose one element from Standards 2, 3, and 4.

Element II.d. Teachers adapt their teaching for the benefit of students with special needs.

Element III.a. Teachers align their instruction with the *NCSCS*.

Element IV.a. Teachers know the ways in which learning takes place, and they know the appropriate levels of intellectual, physical, and emotional development of their students.

Each element had been identified as a need in Scoring Study 1 report and lesson performance data collected after Scoring Study 1.

Because Concordia Public Schools had strong participation in the pilot, I asked the Concordia Pilot lead to share her facilitator's experience and success thus far during Intervention 2, Webinar 2. Intervention 2 webinar was held March 9th, 2015. The title of the Concordia Pilot Lead's presentation was "Setting up Successful Structures: Best practices for facilitating PD sessions." Following the facilitation delivery highlighting some best practices for facilitating interaction among OCT participants, I identified to the participants that Standards 2d, 3d, and 4a would be our elements of review based on statewide OCT performance data.

I started the review with a basic best practice for NCEES, reviewing the element description highlighted in gray. I reminded participants to rate on just what they see in

the video, unlike an authentic observation where the evaluator considers what they see during the observation and what they know, either from other visits or artifacts. I identified descriptors related to each of the three standards that are difficult to observe. I also provided evidences for each element and reviewed questions to consider asking a teacher to prompt their reflection as it relates to the specific element. I noted to the participants that knowing the 17 observable elements within the teacher evaluation rubric would help evaluators look for just the descriptors that can be identified during an observation. There are 25 elements within the NCPTS leaving eight non-observable elements to be evidenced by the teachers. The webinar wrapped up with my contact information and a prompt to be looking for additional communication to support the pilot process.

Follow-up support for webinar participants. Following Intervention 2 (Webinar 2), I forwarded two resources for participants to use as support resources during the remainder of the pilot. The first resource was a questions document that I constructed called “Questions for Post Observation Conferences and Summative Evaluations.” The questions resource contains prompting questions to use during a post conference observation or a summative evaluation. Questions for each element within the North Carolina Professional Teacher Standards is included within the document. I also sent participants an Evidence Document. This document consists of examples of evidences created by principals across North Carolina for principals in North Carolina. I facilitated the data collection process through the eight Principals READY meetings held across each region of the state in 2013. The Evidence list is not exhaustive; it serves as a

reference and a good place for an LEA or charter school to start building a local tool of evidences that is relative to their strategic goals for improvement.

Participant frustration feedback. Informal feedback collected from participants during the webinar communicated a frustration caused by multiple priorities. They felt overwhelmed by the multiple responsibilities they have in addition to trying to complete the OCT lessons. I hoped that hearing about how Concordia had developed a specific schedule for all principals to meet together to complete the OCT lessons had been useful for those participants who were attempting the training totally on their own.

On March 23rd, I produced a report titled *OCT Participation Update* with all 32 LEAs represented. Even if an LEA had just one individual participant, this report identified the LEA represented by the one participant. The report identified the number of participants who completed Scoring Study 1, the number of participants who had completed at least one full observation lesson, and the number of participants who completed at least one element-specific lesson. Table 4 shows the number and percentage of participation.

Table 4

2014–15 OCT Participation as of March 23, 2015

OCT Components	Number of participants who completed	Percent
Scoring Study 1	128	35%
One full observation lesson	59	16%
Element specific lesson	74	21%

Note. Adapted from report titled *OCT Participation Update*

LEAs with the most activity within the OCT were Chatham County, Columbus County, Lincoln County, Newton-Conover City Schools, and Pamlico County. I suspect that the participation in these districts are the result of each having a district facilitator; however, the level of facilitation was varied across each district. Overall participation in the OCT Pilot remained low. With only 35% of principals who had agreed to participate in the online program actually engaging in even one component of the training, I was concerned for the successful implementation of the OCT Pilot.

Intervention Cycle 3: Increase Participation

Based on participation data, I decided to focus on increasing participation. I created a checklist with links to all OCT resources and materials. I sent the checklist to all registered users to get up to date with their participation. The OCT checklist included six steps—three optional and three required.

1. OCT Consent Form (required)
2. OCT Survey 1 (optional)
3. Webinar 1: Kick-off for OCT (optional)
4. Scoring Study 1 (required)
5. Complete half of the 17 lessons (required)
6. Webinar 2 (optional)

Participation in the OCT training required a Consent Form to be completed before beginning the process. Direct links to recorded webinars and links to the PowerPoint presentations used during the webinars was provided so participants could watch the two optional webinars. Links to the OCT Participation Guide and the Observation/Evaluation rubric were also provided on the checklist under Step 3 Kick-off webinar. These

resources had already been sent out to participants at the beginning of the pilot, but I included them again within the checklist for convenience and easy access. Resources provided with Step 6, Webinar 2 included a Questions Document that I created to guide questions when conducting post observation conferences and Summative Evaluations. Specific directions for logging into Bloomboard and completing Scoring Study 1 and element-specific lessons were provided. All resources were linked to an OCT Wiki-page I created just to house this process with all needed links. The wiki page served as a one stop shop for all things OCT.

I emailed this Checklist and a link to the wiki on April 1st, 2015. During the week following the email, only two districts—Chatham County and Newton-Conover City Schools—were active in the system. Participants from both LEAs completed lessons and the Newton-Conover pilot participants were working on completing Scoring Study 2. Both districts that responded to the checklist were being facilitated by a district leader serving as the OCT lead. Efforts to increase participation were unsuccessful. I suspect that the checklist arrived at a time of year when principals are busy with planning testing and end of the year programs and happenings. I also think that if a principal had not started the OCT before April 1st, they would be hesitant to begin a 20-hour training at such a late date. Rather than wait until April 1st, I plan to provide the OCT checklist at the beginning of the training as part of the Participants Guide.

Newton-Conover City Schools (NCCS) administered the second scoring study early (March 25th), and all 12 participants in NCCS completed the scoring study by April 8th. NCCS requested early access to the final scoring study in order to fit it into their bimonthly principal meeting schedule. They embedded OCT resources into their

meetings and provided an environment for participants to engage in collaborative conversations about teacher evaluation and the NCEES rubric. Because of NCCS's facilitated approach to the pilot, we highlighted their efforts in a case study separate from the whole group of OCT participants. NCCS participants' performance on percent target agreement, percent discrepant, and scoring bias improved from Scoring Study 1 to Scoring Study 2. Additional findings from NCCS will be shared in the focused case study.

On May 15th, 2015, I provided an updated report to leadership at DPI outlining some preliminary data for consideration moving forward. Four hundred fifty-seven users from 32 LEAs (LEA groups and individual participants) were given access to the OCT. As of May 14, 136 users had completed at least one task in the system and 128 participants had completed the first scoring study. A total of 1,141 lessons had been completed by 96 participants. Sixty-nine participants had completed at least one full observation lesson and 83 participants had completed at least one element-specific lesson. An OCT participation report updated through May 19th, 2015 reported 42 participants had completed Scoring Study 2.

Data from SS2. Sixty-two participants completed Scoring Study 2 before it closed on May 25, 2015. Empirical Education provided NCDPI with an NC Scoring Study 2 Aggregated Report following the same format as the report provided for NC Scoring Study 1 Aggregated Report. Video V181 was used for Scoring Study 2. The report included a distribution of modal and target scores by video. Based on the data, I identified six elements that were most frequently scored incorrectly. The most frequently incorrectly scored elements were:

1. Standard 1a Teachers lead in the classroom;
2. Standard 2c Teachers treat students as individuals;
3. Standard 2d Teachers adapt their teaching for the benefit of students with special needs;
4. Standard 3b Teachers know the content appropriate to their teaching specialty;
5. Standard 3d Teachers make instruction relevant to students; and
6. Standard 4a Teachers know the ways in which learning takes place, and they know the appropriate levels of intellectual, physical, social, and emotional development of their students.

Four of the six elements most frequently scored incorrectly were identified in both Scoring Study 1 and Scoring Study 2. They are 2c, 2d, 3d, and 4a.

Just like Scoring Study 1, Scoring Study 2 provides Distribution of Ratings by Domain. Distribution of Ratings by Domain identified elements with closely related rating scores.

1. Standard 2c Teachers treat students as individuals;
2. Standard 2d Teachers adapt their teaching for the benefit of students with special needs;
3. Standard 3b Teachers know the content appropriate to their teaching specialty;
4. Standard 3d Teachers make instruction relevant to students; and
5. Standard 4a Teachers know the ways in which learning takes place, and they know the appropriate levels of intellectual, physical, social, and emotional development of their students.

Only one element, Standard 3d: Teachers make instruction relevant to students, was identified in Scoring Study 1 and Scoring Study 2 as strongly disagreed upon answers to the target scores.

Standard 2b: Teachers embrace diversity in the school community and in the world challenged participants during their scoring exercise. Observer Agreement with Target Scores by Domain data identified Standard 2b as a standalone element with 34% discrepancy. The next percentage discrepant was much less at 16% for Standard 4a and 10% for Standard 4g. This means a large percentage of participants thought the target score was not correct. This information supports revisiting the master rating notes for further justification of the rating for Standard 2b. A possible change in target score may be necessary.

Improvement in Participant Performance

Percent target agreement was determined for each OCT participant who completed Scoring Study 1 and Scoring Study 2. To do this, we calculated the percent of the 17 scores that agreed exactly with the target scores. For example, if 8 out of the 17 scores provided by an observer matched the target scores, he/she has a 47% target agreement. The improvement data is determined by the mean percent target agreement, the average percent target agreement across all of the Oct participants that completed the scoring study.

A score is considered discrepant if it is off by 2 or more performance levels from the target scores. For example, if the participant input a 2 and the target score was a 4, the input would be considered discrepant. For each participant completing the scoring study, the percent target discrepant is simply the percentage of his/her scores that were

off by 2 or more levels from the target score. The mean percent target discrepant shown is simply the average percent target discrepant across all of the participants that completed the scoring study.

Improvement in Scoring Accuracy

Participant scoring accuracy improved between Scoring Study 1 and Scoring Study 2. Figure 8 displays the percentages of all participants' agreement with target scores. This information includes data from participants who completed one or both of the scoring studies. Scoring Study 1 identified exact agreement of target scores as 40%. Scoring Study 2 identified exact agreement of target scores as 50%. This shows a 10% increase of exact agreement from Scoring Study 1 to Scoring Study 2. In addition, scores either on target or directly adjacent to the target scores increased from 85% (15% discrepant) on Scoring Study 1 to 94% (6% discrepant) on Scoring Study 2.



Source: Empirical Education (2015), pp. 10–11.

Figure 8. Scoring study comparison: Overall agreement.

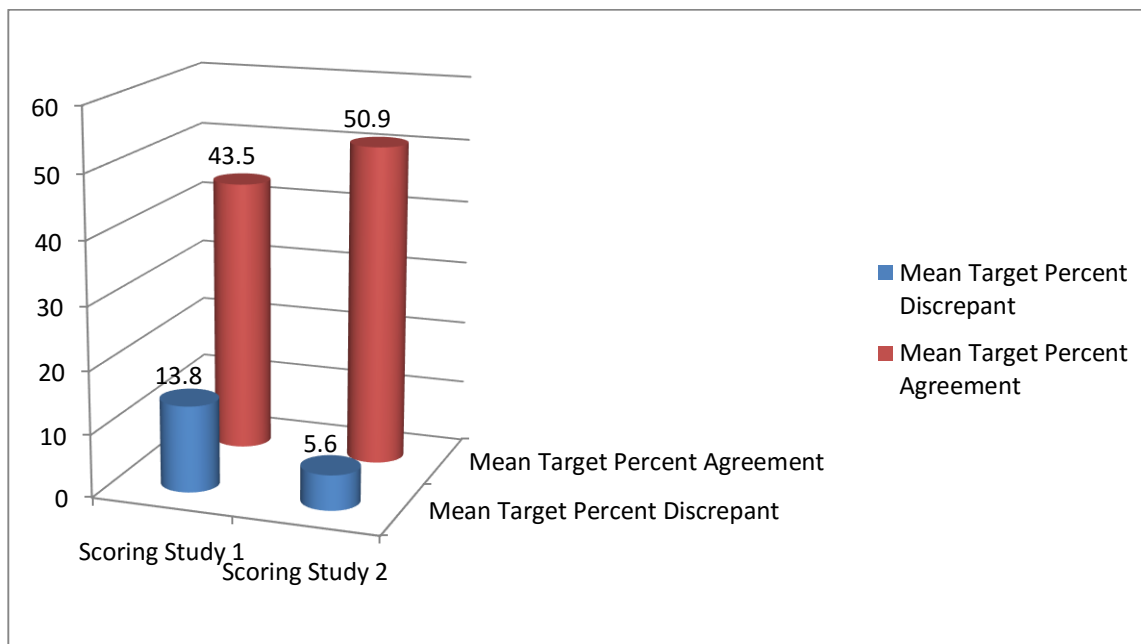
Identification of difficult to rate elements. Participant agreement by individual elements helped identify which elements from Scoring Study 1 and Scoring Study 2 had the highest agreement and the lowest agreement. Figure 9 shows the element agreement graphs for the three elements with the highest agreement and the three elements with the lowest levels of agreement for Scoring Study 1 and Scoring Study 2. Based on the graphs, *Element 2d Teachers adapt their teaching for the benefit of students with special needs* and *Element 3d Teachers make instruction relevant to students* received the most discrepant ratings in both Scoring Study 1 and Scoring Study 2. These elements were both identified as having lowest agreement in both Scoring Study 1 and Scoring Study 2. Lowest agreement of both elements could mean that they are more difficult to rate in video lessons. It could also signify that the language used in the descriptors for each element needs closer examination to truly have an understanding of what teacher behaviors are being rated.

Figure 10 drills down the data to look specifically at the same 60 individuals who completed both Scoring Study 1 and Scoring Study 2. Scoring Study 2 data showed the agreement to target scores as significantly higher than Scoring Study 1. The information in Figure 10 proves that rater accuracy improved from the beginning of the pilot to the end.



Source: Empirical Education (2015), p. 11.

Figure 9. Scoring study comparison: Agreement by element.



Source: Adapted from Empirical Education (2015), p. 11.

Figure 10. Scoring Study 1 & Scoring Study 2 group performance.

Increasing agreement means that scorer accuracy improved. Typically, if there are four or fewer discrete rating levels, the percentage of absolute agreement should both be calculated. Evaluation ratings with better inter-rater agreement are more likely to be a credible source of performance feedback and basis for professional development planning because they are more likely to reflect true strengths and weaknesses rather than a rater's opinion on good educator practice. If agreement on one standard or dimension is consistently low, a revision of the rubric wording or more training on that particular rubric is likely to be needed.

Rater discrepancy identifies the ratings that are more than 1 rating off from the target rating. Rater discrepancy can be helpful when determining whether there were specific domains observers had trouble scoring during an observation. The explanation of the domains may need to be clarified, or additional training on domains with low agreement may be needed. Observers share a common misconception regarding a particular sort of lessons. For example, a group of Observers may tend to give inflated scores to lessons in which students work in cooperative groups while the rest of the group does not. These misconceptions can be addressed in training or with revisions to scoring guidelines. For the OCT, rater discrepancy decreased, telling me that observers did improve.

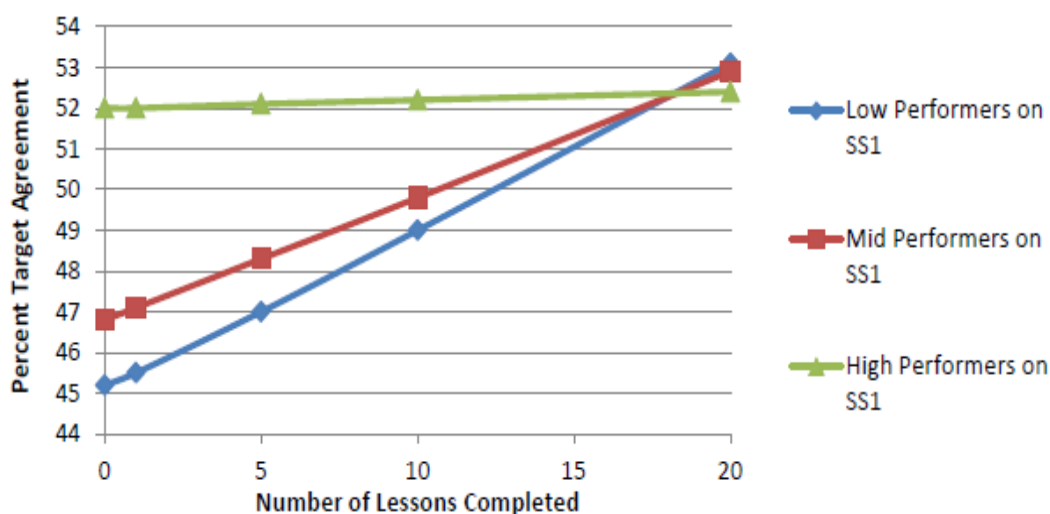
Impact of Number of Lessons Completed

A direct correlation was found between the number of lessons completed by the participants and their improved scoring accuracy.

Regression analysis was used to measure the strength of the association between the number of Lessons completed between Scoring Study 1 and Scoring Study 2 (based on timestamps in Observation Engine) and the two measurements of

performance: percent target agreement and percent target discrepant. (Empirical Education, 2015, p. 13)

Figure 11 shows regression results that participants who performed low or mid-range on Scoring Study 1 had an increased percent target agreement the more lessons they completed. Higher performing participants, based on Scoring Study 1, did not show as strong a correlation. We see a similar performance pattern with high-performing students in school. They grow less because they have less room for improvement.



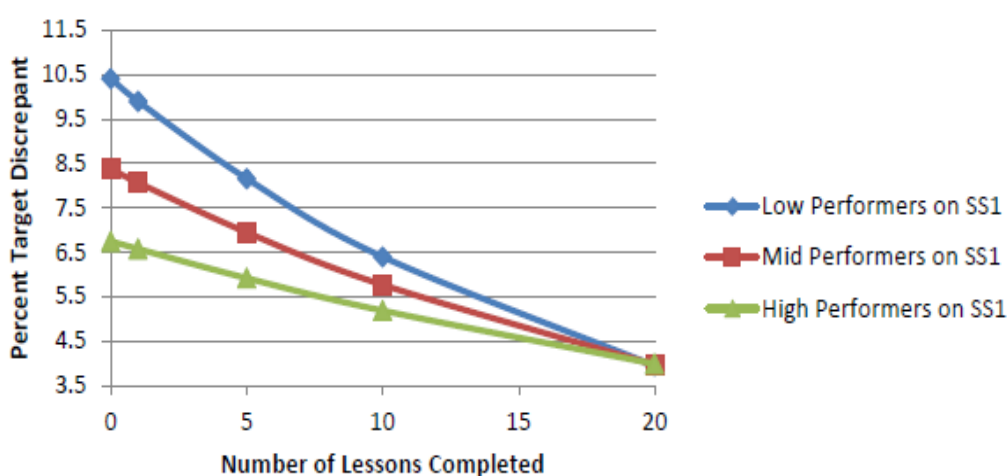
Note: These exact data points are theoretical and not representative of any individual observer. Performance trend lines indicate performance rankings on SS1 (i.e. "pre-test" performance). Low = 25th percentile, Mid = 50th percentile, High = 75th percentile.

Source: Empirical Education (2015), p. 14.

Figure 11. Regression model: Effect of lessons on target agreement.

Target discrepancy supports the more lessons completed, the lower the percent of discrepant scores by all performance levels of participants, based on Scoring Study 1 (see Figure 12). Much like the data on percent target agreement, lesson completion had the greatest positive effect on the lowest performing participants, again based on Scoring Study 1. The blue line has the steepest decline showing the most change.

During the OCT, I recognized that participation in the OCT was key to the project's success. I recognized that the more participation we had in the project, the more data pieces we would have to inform our future implementation. I did not realize during the pilot that the number of lessons completed would have the most significant effect on performance. This is definitely one of the reasons developers and implementers need to work hard at achieving high levels of participation in an online training program.



Note: These exact data points are theoretical and not representative of any individual observer. Performance trend lines indicate performance rankings on SS1 (i.e. "pre-test" performance). Low = 25th percentile, Mid = 50th percentile, High = 75th percentile.

Source: Empirical Education (2015), p. 14.

Figure 12. Regression model: Effect of lessons on target discrepancy.

Summary of Improvement Evidence

Scoring improved between Scoring Study 1 and Scoring Study 2. Percent target agreement increased and percent target discrepant decreased.

- **The number of lessons completed by the observer increased the likelihood that scoring would improve.**

CHAPTER FIVE: IMPACT, DISCUSSION, AND STEPS

Observation Calibration Training (OCT) tested ways the North Carolina Department of Public Instruction can develop and implement a process for improving rater agreement performance on Standards 1 through 5 of the North Carolina Professional Teacher Standards. The PDSA improvement cycle was embedded in the 2014–15 OCT pilot and will continue as a program practice moving forward. By measuring progress of participants after each successive intervention, this project identified ways to enhance the delivery and effectiveness of the online training. For the 6,000 school leaders who will have access to the training, the improvements should increase the capacity to improve the quality and value of educator evaluations and foster improved teaching and overall student achievement. Discussion of the impact of the project and next steps are the focus of this final chapter.

Continuation of Intervention for North Carolina

The results of this project informed the 2015–16 implementation of the OCT statewide plan for improving rater agreement performance on Standards 1 through 5 of the North Carolina Professional Teacher Standards. Webinars and resources during the OCT were reported as beneficial by participants. These and other strategies implemented during the first year will be sustained and improved upon.

Recommended Changes to the Intervention Process

Data and participant feedback clearly supported the continuation of the OCT. An online survey was sent out to all 138 participants, and 42 participants (30%) completed the survey. Of those respondents, 79% reported that they very much or somewhat felt that they improved their application of the NCEES rubric between Scoring Study 1 and

Scoring Study 2. However, pilot participation and performance reports, along with specific feedback from participants, identified needed adjustments within the online platform, Observation Engine, and within the OCT content. Recommended changes include:

- Justify target scores
- Expand video content
- Increase participation and facilitation
- Adjust technical aspects of the program

The target score (rating) justifications within Observation Engine should be expanded for some elements. This is understandable, given observations from scholars who say, “The ongoing challenge for many states is developing an accurate understanding of different level of teaching quality that is shared by all educators” (Gandha & Baxter, 2015, p. 7). During the pilot, participants—individually, and as part of facilitated groups—challenged some the ratings of the videos. Upon further investigation, ratings were found to be accurate; however, they were not fully justified within the target score justification which appears immediately following an incorrect scoring response during a lesson in Observation Engine. Further work and dialogue with participants in this area should result in greater understanding of the rationale for target scores.

The aggregated scoring study data, including a Distribution of Ratings by Domain report and Observer Agreement with Target Scores, indicates that NCDPI and BloomBoard should review video evidence to ensure that the target scores are accurate and justifications are specific and thorough for Standards 2b, 2d, and 3d. In addition,

NCDPI should also look further into these standards to see if there could be a possible misconception that could be clarified through professional development and resource development. Observer bias or misinterpretation may also weigh in as part of the discrepancy.

An additional recommendation stemmed from the overall video library within the OCT. Participants suggested the video library be expanded to include teachers of more varying quality. The majority of videos used displayed teacher behaviors with target scores of “proficient” or less, which includes “developing” and “not demonstrated.” Participants wanted examples of “accomplished” and “distinguished” teacher behaviors. Participant feedback also noted that the teachers in the videos were not from North Carolina. By expanding the video library to include varying quality of teacher behaviors as well as North Carolina teachers modeling successful 21st century teaching strategies, the lessons would better reflect the culture and instructional ranges of North Carolina classrooms

Impact of Collaborative Participation

The OCT platform provides the opportunity for administrators to improve their ratings through individual participation and group facilitated participation. In response to administrators’ requests across the state to the Educator Effectiveness division, the OCT was developed to provide a self-paced, self-directed professional development with flexible time of participation. However, feedback from participants clearly communicated the value they found in the collaborative conversations around content within the OCT. Some LEAs hosted facilitated collaboration for their LEA participants. A case study in Concordia Schools described used multiple activities to support their

heavily scheduled and facilitated participation in the OCT. The Concordia Schools case study was a partner project to this disquisition project.

Although the OCT was designed to provide individualized learning, Newton-Conover City Schools leveraged the OCT resources to create a customized, intensive NCEES professional development program that benefited their participants. (Mullins, 2016). Participants concentrated much of their feedback around the OCT providing a shared context, allowing for understanding of varying perspectives of the video. The group particularly liked the opportunity to “dig deeply” into the North Carolina Professional Teacher Standards (NCPTS) as a collaborative group, paying close attention to the wording and intent of the element within the standard. Concordia Schools’ leadership and participants were very clear that the collaborative time was more beneficial to them than completing the OCT video lessons independently. However, they did feel that they had more experience than most using the NCEES, and that new evaluators would definitely benefit from the independent experience that is self-paced.

It is also likely that encouraging more collaborative group activities locally would maximize the benefit of the tool and increase participation. The final report for the OCT revealed that the more lessons participants completed, the more growth they had during the OCT. This information supports the need to encourage participation and engagement. Through this project, I discovered that participation was greater within an LEA group. This could have been due to the accountability of training completion to someone within the LEA chain of command, or a sparked interest based on colleague conversations about the OCT content. Regardless, feedback from individuals participating in a facilitated setting communicated the collaborative conversations were important and beneficial.

The discussion of individual elements, including descriptors and rating variations, provided valuable insight from a shared experience hosted within Observation Engine.

Qualitative data from this project support the recommendation of a state-provided facilitator guide for LEAs and charter school leaders to use when participating as a group in the OCT. The facilitator's guide would provide optional activities, resources, and information to enrich the collective experience of observers. North Carolina listened to the recommendation and created an OCT wiki page that houses support for the OCT. Within the wiki is a Facilitator's Guide created for LEA and charter school OCT Local Leads (see Appendix A).

In addition to the Facilitators Guide, district leaders will be identified by the OCT project managers. Additional support and communication will go to the district leaders in an attempt to encourage the collaboration outlined in the Facilitators Guide. Additional webinars will be scheduled to support district leaders' use of the Facilitators Guide and how to access participation and performance reports of their LEA participants.

Finally, the last change recommended for the OCT includes the technical use of the Observation Engine. Participants requested a technical adjustment within the tool to allow participants to switch between elements within the full video scoring activities and full video lessons. Rather than rate a video beginning with the first observable element, participants would prefer to have the flexibility within the tool to begin with whatever element they chose and progress and move freely between observables as long as they completed rating all the elements.

Additional State Recommendations

North Carolina should use data from the OCT to inform their support of administrators across the state with standardized training materials. The majority of states within the Southern Regional Education Board (SREB) believe that communicating “key messages” to all teacher evaluators across the state provides “opportunities to address major misconceptions that hinder system implementation” (Gandha & Baxter, 2015, p. 8). Standardized training materials are more cost efficient than face-to-face deliveries from state consultants. NC support for administrators is provided regionally by consultants from the Raleigh home office and by consultants who reside in the field. Providing standardized training materials that include presentation slides, visuals, talking points, and handouts will ensure additional clarification to the process or content. In some cases, these training materials may stand alone or be embedded into other deliveries.

Policy Considerations

Currently, there are no policies regarding the qualifications or abilities of teacher evaluators in North Carolina. Yet, accuracy and reliability of teacher evaluations would be improved with a certification process for evaluators that entailed completing a basic level of professional development. North Carolina requires teacher evaluators to complete a principal preparation program and hold a principal’s license. However, requiring all teacher evaluators to go beyond licensure to become certified to evaluate would ensure that anyone evaluating teachers has sufficient knowledge to recognize effective and ineffective teacher behaviors and to accurately rate teachers on the professional teacher standards.

Barriers to a certification program include the overwhelming responsibility that principals have to do all that is required within the principalship. No one wants to add more to the plate of this overworked position. Therefore, my recommendation would be to offer an optional certification program that is individual, self-paced, and has flexible scheduling for completion and is funded to provide financial incentives for principals who complete the certification. With more and more emphasis on teacher status based on the five professional teacher standards and *Standard 6, student achievement*, principals want and need additional validity to their rating judgments. Empowering evaluators with an option of certification may be viewed as support of the NCEES rather than an additional mandate for compliance.

Final Thoughts

North Carolina has been at the forefront of innovative practice to support teacher evaluation; however, the art of evaluation within its subjectivity remains a concern. The OCT project has provided insight and clarity to better support the NCEES. However, evaluations and interpersonal relationships leave us in a judgment dilemma in which “Educators are realizing the practice of observing and judging teaching is as complex as teaching itself” (Gandha & Baxter, 2015, p. 10). A continuous cycle of improvement, within the collaborative work with LEAs, has and will continue to inform the state level work to support and ensure educator effectiveness.

REFERENCES

- Antonioni, D., & Park, H. (2001). The relationship between rater affect and three sources of 360-degree feedback ratings. *Journal of Management*, 27(4), 479–495.
- Batton, D., Britt, C., DeNeal, J., & Hales, L. (2012). *NC Teacher Evaluations & Teacher Effectiveness: Exploring the relationship between value-added data and teacher evaluations* (Project 6.4). Retrieved from <http://www.ncpublicschools.org/docs/intern-research/reports/teachereval.pdf>
- Bill & Melinda Gates Foundation. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET Project's three-year study*. Retrieved from http://www.metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf
- BloomBoard + Empirical Education. (n.d.). *BloomBoard + Empirical Education: A partnership built to support North Carolina educators*. Retrieved from <http://nnces.ncdpi.wikispaces.net/file/view/NC-EE-BB.pdf/564567297/NC-EE-BB.pdf>
- Bolman, L., & Deal, T. (2008). *Reframing organizations: Artistry, choice, and leadership* (4th ed.). San Francisco, CA: Jossey-Bass.
- Boundless. (2015). *Flattening hierarchies*. Retrieved from <https://www.boundless.com/management/textbooks/boundless-management-textbook/organizational-structure-2/trends-in-organization-27/flattening-hierarchies-156-3983/>
- Clifford, M., Behrstock-Sherratt, E., & Feters, J. (2012). *The ripple effect: A synthesis of research on principal influence to inform performance evaluation design*. A

quality school leadership issue brief. American Institutes for Research. Retrieved from <http://files.eric.ed.gov/fulltext/ED530748.pdf>

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E. H., & Rothstein, J. (2011).

Getting teacher evaluation right: A background paper for policy makers.

Retrieved from http://iaase.org/Documents/Ctrl_Hyperlink/

Session_30c_GettingTeacherEvaluationRight_uid9102012952462.pdf

Empirical Education. (2015). *Final report on North Carolina's pilot of Observation*

Calibration Training 2014–2015. Palo Alto, CA: Author.

Fetter, J. (2013). High fidelity: Investing in evaluation training. Retrieved from

http://www.gtlcenter.org/sites/default/files/docs/GTL_AskTeam_HighFidelity.pdf

Gandha, T., & Baxter, A. (2015). *Toward trustworthy and transformative classroom*

observations: Progress, challenges, and lessons in SREB states. Atlanta, GA:

Southern Regional Education Board. Retrieved from

http://publications.sreb.org/2015/SREB_COReportOnline.pdf

Graham, M. (2011). *What is inter-rater agreement and how can designers of teacher*

evaluation systems maximize it? Retrieved from <http://cecr.ed.gov/>

[compensation/researchSyntheses/34008_CECR_RS_Inter_Rater_](http://cecr.ed.gov/compensation/researchSyntheses/34008_CECR_RS_Inter_Rater_)

[measurement_508.pdf](http://cecr.ed.gov/compensation/researchSyntheses/34008_CECR_RS_Inter_Rater_measurement_508.pdf)

Graham, M., Milanowski, A., Miller (2012). *Measuring and promoting inter-rater*

agreement of teacher and principal performance ratings. Retrieved from

<http://files.eric.ed.gov/fulltext/ED532068.pdf>

Guskey, T. R. (2000). *Evaluating professional development*. Thousand Oaks, CA:

Corwin.

- Guskey, T. R. (2002). Does it make a difference? Evaluating professional development. *Educational Leadership*, 59(6), 45–51.
- Guskey, T. R. (2009). Closing the knowledge gap on effective professional development. *Educational Horizons*, 87(7), 224–233.
- Ho, A., & Kane, T. (2013). *The reliability of classroom observations by school personnel*. Retrieved from http://www.metproject.org/downloads/MET_Reliability%20of%20Classroom%20Observations_Research%20Paper.pdf
- Langley, G. J., Moen, R. D., Nolan, K. M., Nolan, T. W., Norman, C. L., & Provost, L. P. (2009). *The improvement guide: A practical approach to enhancing organizational performance* (2nd ed.). San Francisco, CA: Jossey-Bass.
- Levy, P., & Williams, J. (2004). The social context of performance appraisal: A review and framework for the future. *Journal of Management*, 30, 881–905.
- McClellan, C., Atkinson, M., & Danielson, C. (2012). *Teacher evaluation training and certification: Lessons learned from the measures of effective teaching project* [White paper]. Retrieved from <http://education.ky.gov/teachers/pges/geninfo/documents/teacher%20evaluator%20training%20and%20certification.pdf>
- McGuinn, P. (2012). The state of teacher evaluation reform: State education agency capacity and the implementation of new teacher-evaluation systems. Retrieved from http://www.americanprogress.org/wp-content/uploads/2012/11/McGuinn_TheStateofEvaluation-1.pdf
- Moen, R. (2015). *Foundation and history of the PDSA Cycle*. Retrieved from https://www.deming.org/sites/default/files/pdf/2015/PDSA_History_Ron_Moen.pdf

- NGA Center for Best Practices. (2011). *Preparing educators to evaluate teachers*. Issue Brief. Retrieved from <http://www.nga.org/files/live/sites/NGA/files/pdf/1110PRINCIPALEVALUATION.PDF>
- North Carolina Department of Public Instruction. (2015a). *2014–15 Performance and Growth of North Carolina Public Schools*. Retrieved from <http://www.dpi.state.nc.us/docs/accountability/reporting/exsumm15.pdf>
- North Carolina Department of Public Instruction. (2015b, February). *Highlights of the North Carolina Public School Budget*. Retrieved from <http://www.ncpublicschools.org/docs/fbs/resources/data/highlights/2015highlights.pdf>
- P21. (2015). *Partnership for 21st Century Learning*. Washington, DC: Author. Retrieved from <http://www.p21.org/index.php>
- Performance review calibration-building an honest appraisal. (n.d.). Retrieved from http://www.successfactors.com/en_us/lp/articles/performance-review-calibration.html
- Reform Support Network. (2013). *Promoting Evaluation Rating Accuracy: Strategic Options for States*. Retrieved from <http://www2.ed.gov/about/inits/ed/implementation-support-unit/tech-assist/evaluation-rating-accuracy.pdf>
- Robinson, V. (2011). *Student-centered leadership* (Vol. 15). New York, NY: John Wiley & Sons.
- Simmons, K., & Mullins, H. (2013). *Adaption of “The Ripple Effect.”* Retrieved from <http://edlstudio.wikispaces.com/file/view/ripple.png/527174698/ripple.png>

- Southern Regional Education Board. (2014). *North Carolina: Taking stock and pushing forward*. 2014 State Progress Report. Atlanta, GA: Author. Retrieved from http://publications.sreb.org/2014/NC_2014_Goals.pdf
- Taylor, E. S., & Tyler, J. H. (2012, Fall 2012). Can teacher evaluation improve teaching? *Education Next*, 12. Retrieved from <http://educationnext.org/>
- Tomberlin, T. (2014). READY principals. *NCEES*. Retrieved from <http://nnces.ncdpi.wikispaces.net/READY+Principals+Spring+2014>
- U.S. Department of Education. (2015). *Common Core of Data (CCD): Local Education Agency (School District) Universe Survey Data, 2012-13 v.1a*. Retrieved from <http://nces.ed.gov/ccd/pubagency.asp>
- Wagner, L. (2013). *Teachers worried and confused over new contract system*. Retrieved from <http://www.ncpolicywatch.com/2013/11/20/teachers-worried-and-confused-over-new-contract-system/print/>
- Waliga, H. (2015). *New report ranks North Carolina teachers' pay*. ABC, Inc. Retrieved from <http://abc11.com/education/new-report-ranks-north-carolina-teachers-pay-563728/>
- Yoon, K. S., Duncan, T., Lee, S. W., Scarloss, B., & Shapley, K. L. (2007). *Reviewing the evidence on how teacher professional development affects student achievement*. Retrieved from REL Southwest Regional Educational Laboratory at Edvance Research, Inc. website: http://ies.ed.gov/ncee/edlabs/regions/southwest/pdf/rel_2007033_sum.pdf

APPENDIX: SUPPORTING ARTIFACTS

Following is a list of links to the supporting artifacts:

- [Facilitator Guide](#)
- [OCT Participant Guide](#)
- [Observation/Evaluation Rubric](#)
- [Observation/Evaluation Rubric—Fillable](#)
- [OCT Kick-off Webinar Presentation](#)
- [Questions for Post-Observation Conferences and Summative Evaluation](#)
- [Evidences for Professional Teacher Standards 1–5](#)
- [NC Scoring Study 1 Report](#)
- [NC Scoring Study 2 Report](#)
- [NC Aggregated report for 13 School Districts](#)
- [NC OCT Final Report](#)
- [Newton-Conover Final Report](#)