RESEARCHREPORT

AN EMPIRICAL EDUCATION REPORT

# Final Report on North Carolina's Pilot of Observation Calibration Training 2014-2015

Empirical Education Inc.

*June 30, 2015*

## Table of Contents

Reference this report: Empirical Education Inc. (2015). *Final Report on North Carolina's Pilot of Observation Calibration Training 2014-2015*. Palo Alto, CA: Empirical Education Inc.

# Background

## OVERVIEW OF THE OBSERVATION CALIBRATION TRAINING

The Observation Calibration Training (OCT) provides North Carolina school districts with access to a suite of calibration and training activities for school administrators across the state to improve the accuracy and reliability of teacher evaluations. The online platform combines BloomBoard's professional development resources with Empirical Education's observer training and calibration tool, Observation Engine™. The BloomBoard resource library contains thousands of streaming videos, eBooks, articles, presentations, and self-paced courses. For the OCT, a selection of full-length classroom videos and short video clips in Observation Engine were master scored by a team of experts using the North Carolina Educator Evaluation System (NCEES; see appendix for full rubric). These videos were then made available to observers as rater calibration events called *Scoring Studies* and element-specific learning exercises called *Lessons*. Observation Engine Scoring Studies and Lessons are described below:

- *Scoring Studies:* Scoring Studies help build consensus and inter-rater reliability among a group of evaluators. A study assigns a video (or set of videos) 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.
- *Lessons:* Lessons provide targeted, self-paced online learning activities for evaluators and/or teaching staff. 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 local education agencies (LEAs) to host collaborative conversations to improve instructional leadership skills.

## OCT PILOT

During the 2014/15 school year, the North Carolina Department of Public Instruction (NCDPI), BloomBoard, and Empirical Education initiated a pilot implementation of the OCT. The purpose of the first year pilot was to evaluate the effectiveness of the OCT resources (in particular, the Observation Engine resources) and gather feedback from administrators. By providing Scoring Studies at the beginning and at the end of the pilot, it was possible to measure improvement as a result of the activities during the pilot.

The table below shows the activities included in the pilot. On November 19, participants were provided with an introduction to the project and available resources 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". Several webinars were offered throughout the pilot period to provide feedback on participation and performance, as well as to offer tips and strategies for increasing calibration and scoring accuracy.

TABLE 1. OCT SCHEDULE OF ACTIVITIES AND TASKS

| Date/completion window | OCT task or activity | Description |
|---|---|---|
| **11/19/14** | OCT kick-off | Live webinar introduction to pilot and OCT platform |
| **11/19/14 – 2/3/14** | Scoring Study 1 (SS1) | One-video observation where participants rated all observable NCEES elements. No immediate feedback was provided. |
| **11/18/14 – 6/15/15 (Continuous)** | Element-specific Lessons | Short video clips focused on one specific NCEES element (17 Lessons available with 2 clips for each observable element). Observers watched and rated the clip and then received immediate feedback on their scores. |
| **11/18/14 – 6/15/15 (Continuous)** | Full-observation Lessons | Longer video focused on all 17 observable elements (2 available in OCT pilot). Observers watched and rated the clip, and then received immediate feedback on their scores. |
| **2/10/14** | OCT webinar: Results of SS1 | Reviewed results of Scoring Study 1 and provided tips for increasing calibration. |
| **3/9/15** | OCT webinar: Setting up successful structures | One LEA shared experience facilitating collaborative approach to OCT. Also, using data from SS1, DPI identified trends from SS1 data and provided coaching tools for principals to use during evaluation conferences. |
| **4/22/15 – 5/23/15** | Scoring Study 2 (SS2) | One-video observation where participants rated all observable NCEES elements |
| **5/15/15 – 6/23/15** | End-of-pilot feedback survey | Online survey eliciting feedback from participants on their experience during the pilot. |
| **6/17/15** | OCT Webinar: Results of SS2 & pilot wrap-up | Compared results of Scoring Study 1 and Scoring Study 2, providing feedback on particularly challenging elements. Also showed results of statistical analyses of performance improvements and effects of completing Lessons on scoring accuracy. |

This report presents participation results, performance outcomes for Scoring Study 2 in comparison to Scoring Study 1, and feedback from a focus group discussion and the end-of-pilot survey. Usage information from BloomBoard's resource library is also provided.

# Results

## PILOT PARTICIPATION

### Overview of Pilot Participation

For the 2014/15 pilot, NCDPI reached out to LEA personnel across the state to elicit participation. Originally, 23 LEAs and individual principals from 9 additional LEAs agreed to participate (see Table 2). In total, 457 principals and other evaluators (observers) were added to the platform. Of the initial 32 LEAs, 12 did not participate (i.e. no observers completed tasks in the platform). By the end of the pilot, 138 observers across 20 LEAs completed at least one task in the system. It should be noted that there are many reasons why users initially added to the platform may not have participated. Over-inclusion of staff members in the original list submitted to NCDPI, personnel changes, and shifts in LEA priorities and resources likely all contributed to participant attrition.

TABLE 2. PARTICIPATING LOCAL EDUCATION AGENCIES

| LEA name | No. of observers initially added | No. of observers completing one or more tasks |
|---|---|---|
| Alexander County Schools | 4 | 1 |
| Alleghany County Schools | 13 | 0 |
| American Renaissance School | 3 | 0 |
| Asheboro City Schools | 3 | 0 |
| Asheville City Schools | 14 | 5 |
| Buncombe County Schools | 1 | 0 |
| Cabarrus County Schools | 9 | 2 |
| Camden County Schools | 7 | 7 |
| Chapel Hill-Carrboro City Schools | 2 | 2 |
| Charlotte-Mecklenburg Schools | 108 | 1 |
| Chatham County Schools | 31 | 13 |
| Clay County Schools | 1 | 1 |
| Columbus County Schools | 26 | 17 |
| Edgecombe County Public Schools | 1 | 0 |
| Elkin City Schools | 1 | 0 |
| Gates County Public Schools | 12 | 8 |
| Guilford Preparatory Academy Charter School | 1 | 0 |
| Jones County Public Schools | 11 | 10 |
| Lincoln County Schools | 59 | 10 |
| Newton-Conover City Schools | 16 | 13 |
| Northampton County Schools | 22 | 8 |
| Pamlico County Schools | 14 | 13 |
| Person County Schools | 6 | 3 |
| Roanoke Rapids Graded School District | 14 | 8 |
| Rockingham County Schools | 55 | 6 |
| Rowan-Salisbury Schools | 1 | 0 |
| Stanly County Schools | 2 | 0 |
| Sugar Creek Charter School | 1 | 0 |
| Tar River Academy | 1 | 0 |
| Union Academy Charter School | 1 | 0 |
| Vance County Schools | 2 | 1 |
| Warren County Schools | 15 | 9 |
| **TOTAL** | **457** | **138** |

## Scoring Study Participation

There were two Scoring Studies administered during this pilot. Scoring Studies are online calibration events where a group of observers first watch and rate a video independently, and then a report is run to compare scores from observers to target scores. The report displays scoring distributions, agreement trends, and performance metrics to facilitate conversations around the scores and observation rubric. For the purposes of measuring improvement during the pilot, Scoring Study 1 (SS1) was considered a "pretest" at the beginning of the pilot, and Scoring Study 2 (SS2) was considered a "posttest" at the end of the pilot.

Table 3 shows Scoring Study participation by LEA. Participation results for SS2 were lower than for SS1.

TABLE 3. SCORING STUDY PARTICIPATION BY LEA

| LEA name | No. of observers completing SS1 | No. of observers completing SS2 |
|---|---|---|
| Alexander County Schools | 1 | 0 |
| Asheville City Schools | 5 | 1 |
| Cabarrus County Schools | 2 | 0 |
| Camden County Schools | 7 | 3 |
| Chapel Hill-Carrboro City Schools | 2 | 0 |
| Charlotte-Mecklenburg Schools | 1 | 1 |
| Chatham County Schools | 12 | 5 |
| Clay County Schools | 0 | 0 |
| Columbus County Schools | 15 | 10 |
| Gates County Public Schools | 8 | 4 |
| Jones County Public Schools | 10 | 3 |
| Lincoln County Schools | 9 | 5 |
| Newton-Conover City Schools | 13 | 12 |
| Northampton County Schools | 8 | 1 |
| Pamlico County Schools | 12 | 12 |
| Person County Schools | 3 | 0 |
| Roanoke Rapids Graded School District | 8 | 4 |
| Rockingham County Schools | 4 | 1 |
| Vance County Schools | 1 | 0 |
| Warren County Schools | 9 | 0 |
| TOTAL | **130** | **62** |

*Note*: Two observers completed SS1 during a second round opportunity and were not included in the initial presentation of results in February 2015. LEAs with no OCT participation are excluded from this table.

## Lesson Participation

There were 19 total Lessons available in the OCT: 17 element-specific Lessons and 2 full-observation Lessons. Participants were asked to complete all available Lessons. The table below shows Lesson participation rates by LEA. Overall, 101 participants across 16 LEAs completed at least one Lesson. Of all observers that participated in the OCT, 37% (51 of 138) completed all available Lessons.

TABLE 4. LESSON PARTICIPATION BY LEA

| LEA name | No. of observers completing at least 1 element-specific Lesson | No. of observers completing at least 1 full-observation Lesson | No. of observers completing all 19 available Lessons |
|---|---|---|---|
| **Alexander County Schools** | 0 | 0 | 0 |
| **Asheville City Schools** | 1 | 1 | 1 |
| **Cabarrus County Schools** | 2 | 1 | 0 |
| **Camden County Schools** | 3 | 4 | 2 |
| **Chapel Hill-Carrboro City Schools** | 0 | 0 | 0 |
| **Charlotte-Mecklenburg Schools** | 0 | 0 | 0 |
| **Chatham County Schools** | 10 | 5 | 4 |
| **Clay County Schools** | 1 | 1 | 0 |
| **Columbus County Schools** | 13 | 14 | 12 |
| **Gates County Public Schools** | 7 | 6 | 3 |
| **Jones County Public Schools** | 6 | 6 | 3 |
| **Lincoln County Schools** | 9 | 9 | 7 |
| **Newton-Conover City Schools** | 12 | 0* | 0* |
| **Northampton County Schools** | 3 | 3 | 2 |
| **Pamlico County Schools** | 13 | 13 | 12 |
| **Person County Schools** | 0 | 2 | 0 |
| **Roanoke Rapids Graded School District** | 3 | 1 | 1 |
| **Rockingham County Schools** | 4 | 5 | 2 |
| **Vance County Schools** | 0 | 0 | 0 |
| **Warren County Schools** | 4 | 5 | 2 |
| **TOTAL** | **91** | **76** | **51** |

* All observers in Newton-Conover City Schools (NCCS) scored one of the full-observation Lessons as a third Scoring Study, hence it did not make sense for these observers to complete all 19 available Lessons. See the NCCS case study report for further information on NCCS's unique OCT implementation.

*Note*: LEAs with no OCT participation are excluded from this table.

Participation rates for SS2 were lower than for SS1. This may be due to SS2 being administered at the end of the year, a busy time for most administrators. That being said, for a completely voluntary project, this level of participation is quite encouraging. Participation in Lessons was greater than participation in SS2 possibly because of the availability of immediate on-screen feedback, which may make Lessons inherently more attractive to observers.

### BloomBoard Resource Library Usage

Although the focus in the 2014/15 OCT pilot was on the Observation Engine resources, the OCT platform did include access to BloomBoard's resource library. The table below shows usage metrics for the resource library by LEA. The first column lists the number of login/usage sessions. Sessions were counted as any period where a user logged in and was actively navigating the system. The second column breaks down the number of resources that were accessed by users (including any Observation Engine resources). The last column outlines the number of unique searches in the Marketplace (using the search bar or other filters). Lincoln County Schools added approximately 45 admin accounts and Person County Schools added approximately 250 teacher accounts this year so other users could explore resources in the BloomBoard marketplace.

TABLE 5. BLOOMBOARD RESOURCE LIBRARY USAGE

| LEA name | Session count | No. of previewed/consumed resources | No. of searches in marketplace |
|---|---|---|---|
| Alexander County Schools | 6 | 0 | 0 |
| Alleghany County Schools | 5 | 0 | 3 |
| American Renaissance School | 1 | 3 | 5 |
| Asheville City Schools | 35 | 7 | 25 |
| Buncombe County Schools | 1 | 0 | 0 |
| Cabarrus County Schools | 18 | 0 | 4 |
| Camden County Schools | 32 | 4 | 14 |
| Chapel Hill-Carrboro City Schools | 7 | 7 | 8 |
| Charlotte-Mecklenburg Schools | 2 | 0 | 0 |
| Chatham County Schools | 73 | 7 | 5 |
| Clay County Schools | 4 | 0 | 0 |
| Columbus County Schools | 142 | 12 | 28 |
| Gates County Public Schools | 45 | 15 | 22 |
| Jones County Public Schools | 52 | 0 | 7 |
| Lincoln County Schools | 141 | 171 | 472 |
| Newton-Conover City Schools | 144 | 33 | 27 |
| Northampton County Schools | 37 | 6 | 19 |
| Pamlico County Schools | 103 | 2 | 12 |
| Person County Schools | 27 | 34 | 87 |
| Roanoke Rapids Graded School District | 38 | 5 | 16 |
| Rockingham County Schools | 32 | 7 | 15 |
| Vance County Schools | 4 | 0 | 0 |
| TOTAL | **1023** | **352** | **802** |

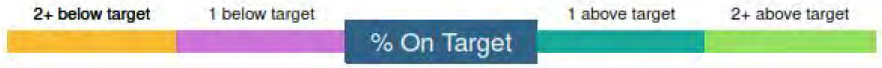
## OBSERVER PERFORMANCE

### Comparison of SS1 and SS2 Based on Reports Generated by Observation Engine

Examining the automatically-generated Observation Engine scoring reports for both Scoring Studies, there appeared to be improvement in scoring accuracy between SS1 and SS2. To interpret these graphs and subsequent analyses, three performance metrics are defined:

- *Percent target agreement*: the percent of an observer's scores that agrees exactly with the target scores
- *Percent target discrepant*: the percent of an observer's scores that disagrees with the target scores by 2 or more performance levels (e.g. when the target score is 2 and the score provided is a 4)
- *Scoring bias*: when an observer has a statistically significant tendency to rate higher or lower than the target score

The figures below that report the results use the graphic convention shown here:



This multicolored bar represents all scores submitted by observers. The length of each colored section represents the portion of scores that fall in that particular category: the longer the section, the higher the percent of total scores. The value reported in the dark blue section of the bar is the percent target agreement. Scores that were one score adjacent to the target score are represented by the purple ("1 below target") and teal ("1 above target") colored sections. The percent target discrepant is reported to the right of each bar and are also represented by the orange ("2+ below target") and lime green ("2+ above target") colored sections. Any scoring bias would be reported as an upwards-facing or downwards-facing arrow next to the bar.

Figure 1 summarizes the agreement to target scores for all observers who completed each Scoring Study. You can see that in SS2, exact agreement to target scores was higher by 10%. In addition, 94% of all scores in SS2 were either on target or directly adjacent to the target scores (as opposed to 85% for SS1). There was no significant bias towards rating higher or lower than target scores in either Scoring Study.

FIGURE 1. SCORING STUDY COMPARISON: OVERALL AGREEMENT

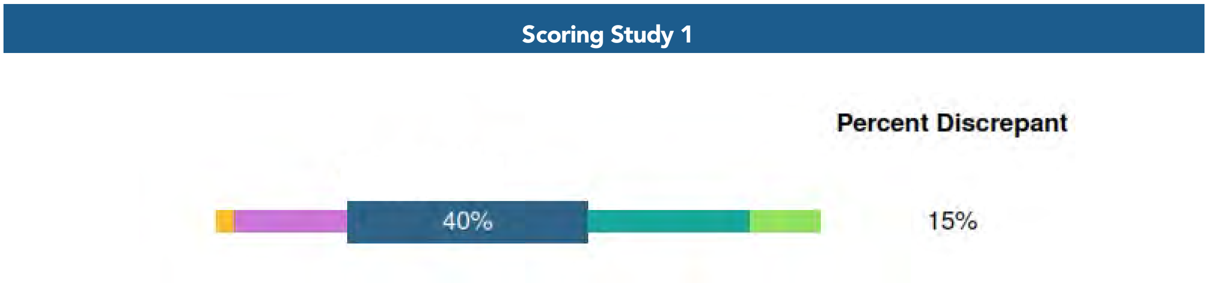
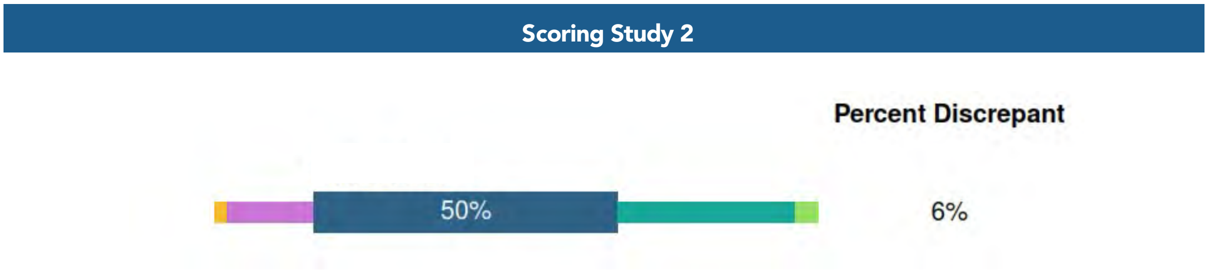
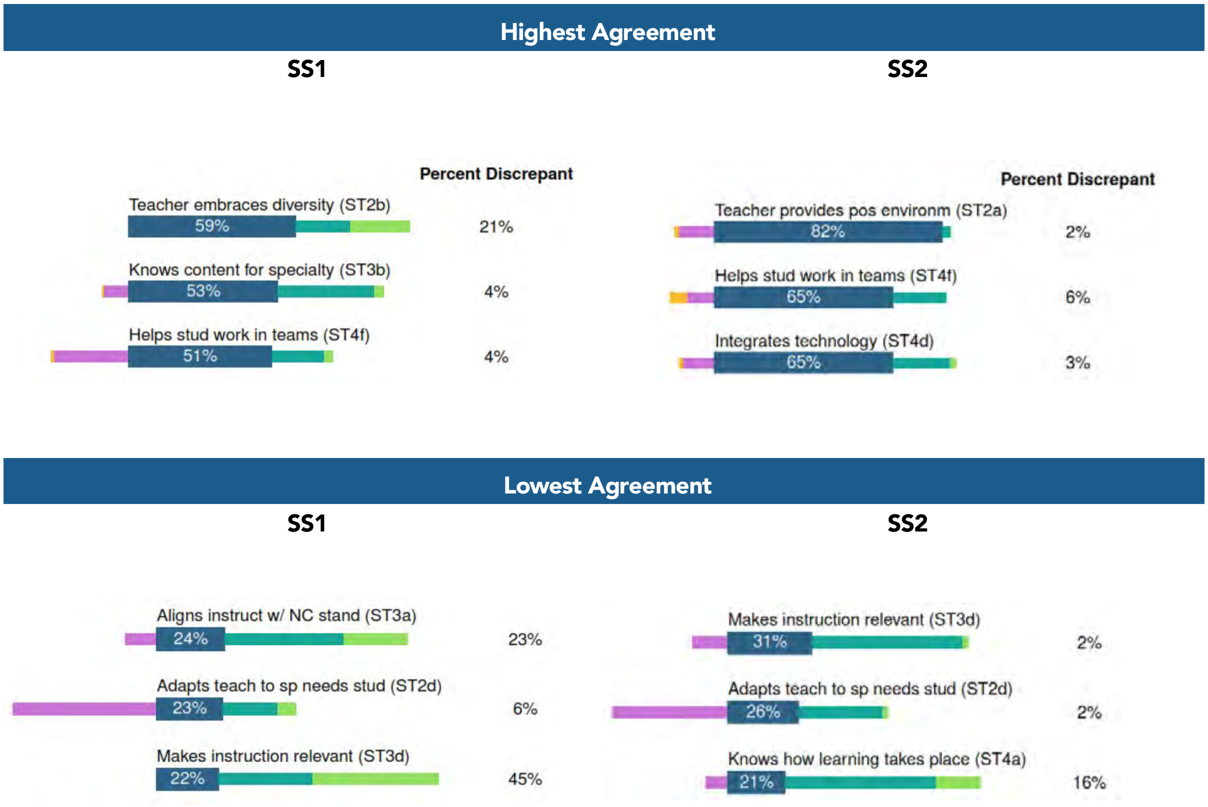


FIGURE 1. SCORING STUDY COMPARISON: OVERALL AGREEMENT



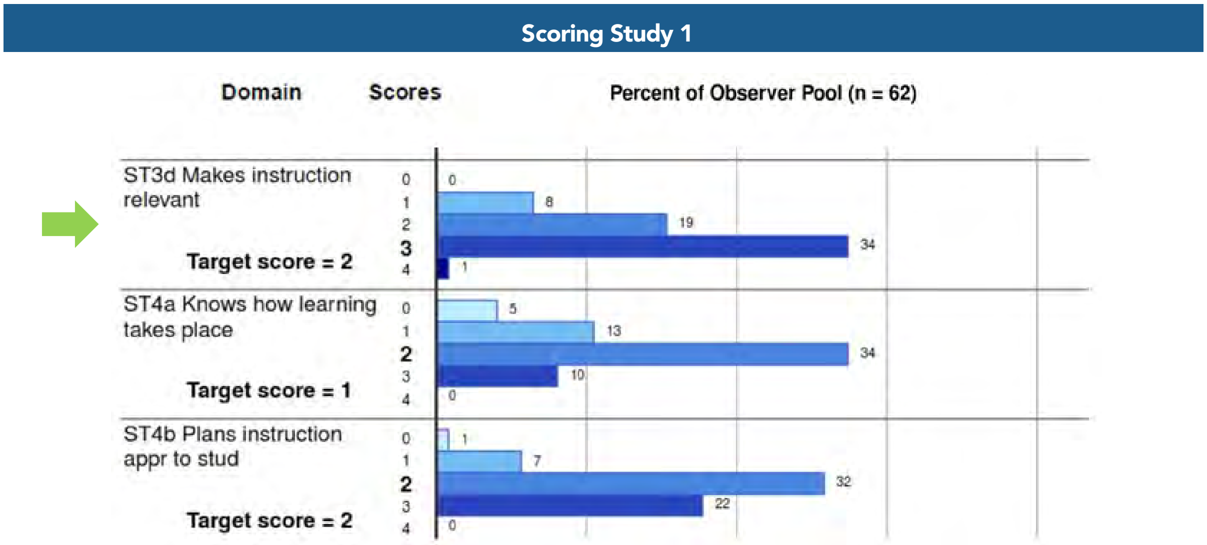
The Scoring Study reports also show agreement by individual NCEES element. Figure 2 shows the agreement graphs for the three elements with the highest and lowest levels of agreement for each Scoring Study. Examining these graphs show that there is some overlap between SS1 and SS2, particularly for the elements with the lowest levels of agreement. Elements 2d and 3d were challenging for observers in both Scoring Studies. This could mean that these elements are particularly difficult to rate in a video observation context. This could also mean that observers should revisit the language of these elements in the NCEES rubric to clarify any confusion or misinterpretation of the language.

FIGURE 2. SCORING STUDY COMPARISON: AGREEMENT BY ELEMENT



Examining the distribution graphs in the Scoring Study reports provides a useful snapshot of scoring that helps elucidate the nature of disagreement with target scores. Figure 3 below shows the distribution of scores in Scoring Study 2 for element 3d – *Teacher makes instruction relevant*.

FIGURE 3. ANALYZING DISAGREEMENT: SAMPLE SCORING DISTRIBUTION GRAPH



The Observation Engine report of the score distribution for element 3d shows that more than half the observers (34 of 62) thought that the score should have been a 3 rather than a 2. The Observation Engine gives observers the opportunity to examine the justifications for the target score to see why the expert scoring committee believed the score to be a 2, and then to examine their own evidence for the score. The report can also be of value to trainers and other evaluation program personnel in investigating the possibility that observers misinterpreted the video, or that they had developed a bias or general impression of the teacher that affected the score on this particular element. It could also mean that the target score should be re-evaluated.

The use of these reports in some of the pilot districts shows that these kinds of explorations of the scoring data can inspire useful collaborative conversations—around the NCEES rubric and evaluation practices—that can contribute to calibration and rater accuracy.

### Statistical Analysis of Scoring Accuracy Improvement

Although the reports for both Scoring Studies generated by Observation Engine showed higher ratings in SS2, statistical analysis allows us to determine whether that improvement was likely due to chance fluctuation. Table 6 shows the results using data from the 60 observers that completed both Scoring Study 1 and Scoring Study 2. The statistical analysis showed that agreement to target scores was significantly higher in SS2 and discrepancy was lower. There was no significant change in scoring bias as there was very little bias to start with.

TABLE 6. SCORING STUDY 1 & SCORING STUDY 2 GROUP PERFORMANCE

| Metric | Scoring Study 1 | Scoring Study 2 |
|---|---|---|
| **Mean Percent Target Agreement** | 43.5% | 50.9%* |
| **Mean Target Discrepant** | 13.8% | 5.6%** |
| **Average Scoring Bias** | 0.15 | 0.18 |

*Percent Target Agreement:* The percentage of scores that exactly match the target score.

*Percent Target Discrepant*: The percentage of scores that disagree with the target score by two or more performance levels.

*Scoring Bias:* Scorer has a statistically significant bias towards rating either higher or lower than the target score.

*Difference from SS1 results statistically significant at *p*<.05

**Difference from SS1 results statistically significant at *p*<.001

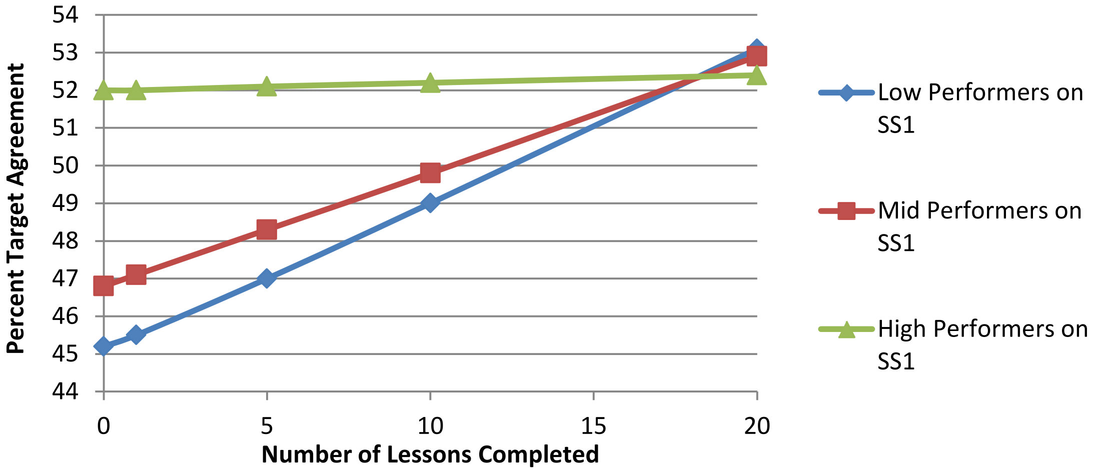
This means that the improvement seen in the reports generated by Observation Engine was not due to chance, and rater accuracy did, in fact, improve from the beginning of the pilot to the end.

### Statistical Analysis: Did Lesson Completion Affect Performance on Scoring Studies?

As previously reported, 19 total Lessons were available to participants in the OCT platform. An important question is whether or not the number of Lessons completed by the observers is associated with improvement between SS1 and SS2. To answer this question, regression analysis was used to measure the strength of the association between the number of Lessons completed between SS1 and SS2 (based on timestamps in Observation Engine) and the two measurements of performance: percent target agreement and percent target discrepant.

Regression results are shown graphically below in Figures 4 and 5. Figure 4 shows that for low- and mid-performing observers on SS1, the more Lessons completed, the higher the percent target agreement. The effect was not as strong for the highest performing observers, which makes some sense since they have less room for improvement between SS1 and SS2.

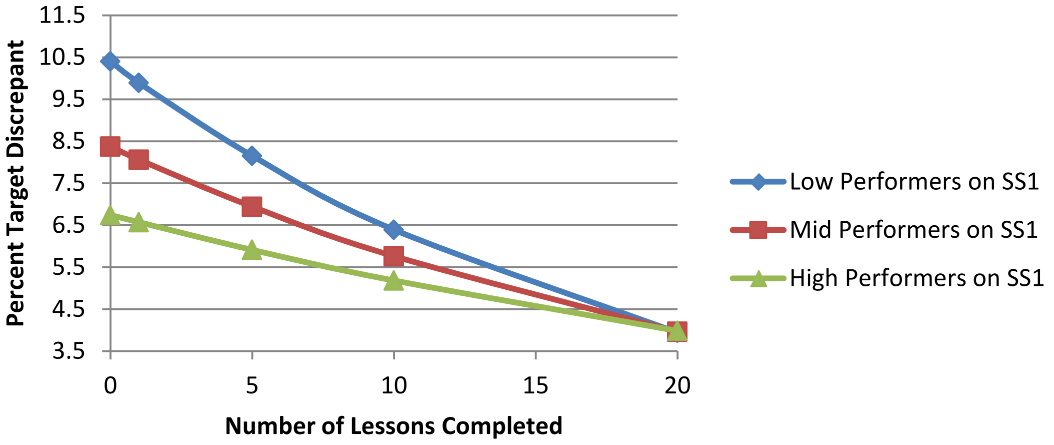
FIGURE 4. REGRESSION MODEL: EFFECT OF LESSONS ON TARGET AGREEMENT



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.

The effect of completing Lessons was even stronger on the percent target discrepant metric. Figure 5 shows that for all SS1 performance levels, the more Lessons completed, the lower the percent of discrepant scores. This means that all observers, regardless of their initial performance on SS1, benefited from completing Lessons. Similar to the effect on percent target agreement, completing Lessons benefited the lowest scoring observers the most (the trend line for the low performers is steepest).

FIGURE 5. REGRESSION MODEL: EFFECT OF LESSONS ON TARGET DISCREPANT



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.

## FINDINGS FROM FOCUS GROUP

On May 19, Empirical Education and BloomBoard hosted a focus group discussion to gather qualitative feedback from participants at Newton-Conover City Schools on their experience with the OCT pilot. Participants shared that the OCT provided an opportunity to "dig deeply" into the NCEES standards as a group. The conversations that ensued around a particular event or instructional practice seen in the video were highly valuable, allowing for understanding of varying perspectives and interpretations. They looked specifically at semantics and interpretations of wording around particular elements in the NCEES framework. These collaborative activities brought far more value to the OCT process than completing the OCT video observations independently. However, the Newton-Conover group acknowledged that they had much more experience with the NCEES and classroom observation than newer evaluators, who would definitely benefit from the OCT's self-paced activities.

Prior to the OCT, Newton-Conover conducted NCEES work by discussing the specific standards and generating look-fors for each element. They were not able to refer to videos of classroom lessons to discuss specific examples. The OCT's video observation capabilities gave them the opportunity to look more in-depth and apply the standards as a group.

The participants expressed that the OCT videos did have some limitations. They believe that some of the target scores were slightly inflated, and that there was not sufficient evidence given for some of the target scores. In addition, it would have been helpful to see some examples of distinguished teaching in the video set, as well as more diverse grade levels and subject areas. They also provided suggestions for improvements to the rating page interface, including the ability to toggle between elements, so that users could see all element-specific information on one page. When asked if they would like to have access to the OCT next year, they said that they would certainly benefit from additional collaborative work with videos they had not yet seen.

In the end, they indicated that they learned a lot from this process. They gained a better understanding of how different evaluators go through the rating process and felt that the collaboration that the OCT encourages improved their understanding of the NCEES framework.

## FEEDBACK FROM PARTICIPANT SURVEY

At the end of the pilot period, an online survey was sent out to all participants. As of June 24, 42 participants completed the survey. Table 7 below shows response distributions for questions related to the introductory material and training. Most respondents (72%) thought that the introductory materials and the kick-off webinar were either useful or very useful. Only one respondent thought these resources were not useful. In comments, it was noted that it would have been helpful to be able to log in prior to the webinar. It was also noted that it would have been helpful to have a more concise schedule: "This was very open-ended, which makes it difficult and easy to put off."

TABLE 7. SURVEY RESULTS: OCT TRAINING AND INTRODUCTORY MATERIAL

| Question | Response | Frequency | Percent |
|---|---|---|---|
| You were sent an introductory email explaining the Observation Calibration Training (OCT) platform and how to login. How useful was it? | Did not use | 1 | 2% |
| | Not useful | 1 | 2% |
| | Somewhat useful | 10 | 24% |
| | Useful | 20 | 48% |
| | Very useful | 10 | 24% |
| How useful was the webinar introducing the OCT? | Did not see webinar | 1 | 2% |
| | Not useful | 1 | 2% |
| | Somewhat useful | 10 | 24% |
| | Useful | 20 | 48% |
| | Very useful | 10 | 24% |
| Note. Not all survey respondents answered every question. | | | |

Table 8 shows survey results related to technical issues with the OCT platform. The vast majority of respondents did not come across any technical issues, but those that did were able to resolve them. Eighty-four percent of respondents thought that the OCT was "easy" or "very easy" to use. Regarding visual and audio quality of the videos, there did not appear to be any major issues. However, some participants did report that there were volume and background noise issues with some of the videos. In addition, it was noted by several respondents that including videos across the entire range of performance levels would have been helpful (i.e. seeing more videos that contain exemplary teaching).

TABLE 8. SURVEY RESULTS: TECHNICAL ISSUES WITH OCT PLATFORM

| Question | Response | Frequency | Percent |
|---|---|---|---|
| Did you experience any technical issues with the OCT platform? | No | 33 | 79% |
| | Yes, but resolved on my own | 3 | 7% |
| | Yes, but resolved with help of another person | 6 | 14% |
| | Yes, but the issue was not resolved to my satisfaction | 0 | 0% |
| How would you rate the OCT's ease of use? | Not easy | 0 | 0% |
| | Somewhat easy | 7 | 17% |
| | Easy | 25 | 60% |
| | Very easy | 10 | 24% |
| Of the videos you watched, how would you rate the quality of the picture? | Poor quality | 0 | 0% |
| | Fair quality | 15 | 37% |
| | Good quality | 25 | 61% |
| | Excellent quality | 1 | 2% |

TABLE 8. SURVEY RESULTS: TECHNICAL ISSUES WITH OCT PLATFORM

| | | | |
|---|---|---|---|
| **Of the videos you watched, how would you rate the quality of the audio?** | Poor quality | 2 | 5% |
| | Fair quality | 19 | 45% |
| | Good quality | 20 | 48% |
| | Excellent quality | 1 | 2% |

Note: Not all survey respondents answered every question.

Table 9 contains feedback on the Scoring Studies. Between SS1 and SS2, 79% of respondents felt that their application of the NCEES rubric improved. The bottom section of the table lists a selection of comments related to Scoring Studies. The participants expressed that they learned a lot from the experience and appreciated having justifications for the scores. The last comment listed provides an interesting suggestion on how the justifications could be elaborated on to be more helpful in understanding disagreement. It was also noted by some observers that it was frustrating to not see immediate feedback on their performance on Scoring Studies. This is of course by design, since observers are meant to come together as a group and review group agreement trends. However, this suggests that in the future, it may be important to ensure that individual observers are given a clear opportunity to review their own results on any OCT task.

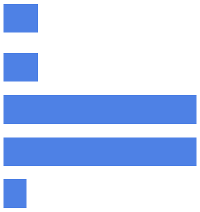
TABLE 9. SURVEY RESULTS: FEEDBACK ON SCORING STUDIES

| Question | Response | Frequency | Percent |
|---|---|---|---|
| **To what extent do you feel your application of the NCEES rubric improved between Scoring Study 1 and Scoring Study 2?** | Very much improved | 4 | 10% |
| | Somewhat improved | 29 | 69% |
| | No different | 8 | 19% |
| | Not applicable | 1 | 2% |
| **Please provide any additional feedback you have regarding your experience with the Scoring Studies, including comparing it to other forms of training you have participated in.** | "I think it was more useful than the initial training that we had especially the feedback on why the scores were different from what we selected."<br>"I think I started to better understanding what we are looking for in our teacher observations through more practice."<br>"I thought it was very informative but some feedback comments I didn't agree with."<br>"It would have been extremely helpful to include what the teacher in each video could/should have done to reach the next level on the rubric.  Often, I felt that my notes were the same as those provided in the "key," but I may have had a different rating; because there was no explanation of what could have been shared with the teacher in the post-conference to move them forward, I was unclear why my rating was "incorrect."" | | |

Note: Not all survey respondents answered every question.

Table 10 presents feedback on OCT Lessons. Although there was some variability in what respondents thought of Lessons, most thought that the target score feedback was either fair or good. The comments in the bottom half of the table help elucidate both strengths and weaknesses of the Lessons. Some of the comments allude to the difficulty inherent in video observation. Only seeing two camera angles and only observing a teacher for a short period of time can make score judgements difficult. It is through extensive practice that observers can improve their scoring accuracy and apply the rubric objectively.

TABLE 10. SURVEY RESULTS: FEEDBACK ON LESSONS

| Question | Response | Frequency | Percent |
|---|---|---|---|
| **In the Lessons you completed, how would you rate the quality and usefulness of the target scores and justifications?** | I did not complete Lessons  | 3 | 7% |
| | Poor | 3 | 7% |
| | Fair | 17 | 40% |
| | Good | 17 | 40% |
| | Very good | 2 | 5% |
| **Please provide any additional feedback you have regarding the use of Lessons, including comparing it to other forms of training you have participated in.** | "I think this type of training would also be beneficial to teachers, it would help give them a better grasp about what we are looking for when we observe their classroom. "<br>"Again, often evidences were really hard to justify due to the camera angles. Even with the artifacts, it was extremely difficult sometimes to say, 'Yes, there it is. I now have this evidence.' Also, in small groups, it was harder to hear exactly what was happening."<br>"I thought that the target scores and justifications were a bit of stretch in some of the lessons. While I agreed with many of the scores and justifications, I thought that a few missed the mark. I think that this really highlighted a flaw of the instrument itself. Many of the elements of each standard are not "observable" during the timeframe of a classroom observation. To accurately assess a teacher in these areas an administrator must have knowledge of what the teacher does over time. "<br>"The overall lesson I learned - be wary of being generous, be "by the book" in terms of the rubric. We must continue to reiterate to teachers that "Proficient" is not on par with 'C-'" | | |

Note: Not all survey respondents answered every question.

## Discussion

The 2014/15 pilot was a small implementation with voluntary participation from evaluators in 20 of the 267[1] North Carolina LEAs. Despite these limitations, the results of this pilot were consistently positive. Observers showed statistically significant improvement on scoring from the start of the pilot to the end. That improvement was directly tied to the extent to which they utilized the available online resources. Newton-Conover City Schools leveraged the OCT resources to create a customized, intensive NCEES professional development program that yielded an enormous amount of benefit for their evaluators. Participants expressed that the OCT platform was easy to use, had few technical issues, and helped them improve their application of the NCEES rubric.

There is certainly room for improvement. The target score justifications could be expanded in some cases, the available library of videos could also be expanded to include teachers of more varying quality, and the schedule of activities could be more concise with more immediate opportunities to review individual performance. It is also likely that encouraging more collaborative group activities either locally or regionally would maximize the benefit of the tool.

This first year pilot of the OCT revealed that the tool provides resources that can help North Carolina administrators become better evaluators.

[1] Data Source: U.S. Department of Education, National Center for Education Statistics, Common Core of Data (CCD), "Local Education Agency (School District) Universe Survey", 2012-13 v.1a.

# Appendix A. North Carolina Educator Evaluation System Rubric

TABLE A1. NCEES TEACHER RUBRIC OBSERVABLE ELEMENTS*

| Standard | Element number | Element description |
|---|---|---|
| **Standard 1: Teachers demonstrate leadership** | 1a | Teachers lead in their classrooms |
| **Standard 2: Teachers establish a respectful environment for a diverse population of students** | 2a | Teachers provide an environment in which each child has a positive, nurturing relationship with caring adults |
| | 2b | Teachers embrace diversity in the school community and in the world |
| | 2c | Teachers treat students as individuals |
| | 2d | Teachers adapt their teaching for the benefit of students with special needs |
| **Standard 3: Teachers know the content they teach** | 3a | Teachers align their instruction with the North Carolina Standard Course of Study |
| | 3b | Teachers know the content appropriate to their teaching specialty |
| | 3c | Teachers recognize the interconnectedness of content areas/disciplines |
| | 3d | Teachers make instruction relevant to students |
| **Standard 4: Teachers facilitate learning for their students** | 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 |
| | 4b | Teachers plan instruction appropriate for their students. |
| | 4c | Teachers use a variety of instructional methods |
| | 4d | Teachers integrate and utilize technology in their instruction |
| | 4e | Teachers help students develop critical-thinking and problem-solving skills |
| | 4f | Teachers help students work in teams and develop leadership qualities |
| | 4g | Teachers communicate effectively |
| | 4h | Teachers use a variety of methods to assess what each student has learned |

*The elements listed in this table only represent the elements that are directly observable through video observation.