National Center for Teacher Effectiveness

# NCTE Webinar Takeaways: Monitoring Rater Reliability

The observation of teacher practice—and the quantification of that practice in evaluators' ratings—have become increasingly important in educator evaluation systems. Validating those ratings as representative of effective teaching practice, and reliable among a school district's team of reviewers, can present major challenges as districts roll out their observation and evaluation systems.

A valid test is one truly measures what it is designed to measure. Reliability means evaluators assign consistent ratings to similar examples of teaching practice. Inter-rater reliability measures the ability of multiple evaluators to agree on ratings of similar examples of teaching practice.

Teachstone Chief Strategy Officer Laurie McCullough says reliability is essential for its research-based protocol to retain its scientific integrity. She explains,

"You need a process to ensure reliability," she says. "It's like learning a new language, with the domains, dimensions, indicators and behavioral markers."'

Inter-rater reliability becomes an even greater issue when there's only one evaluator conducting the observations. Research done by Pearson has found that when one person scores an entire rubric, there tends to be less variability in the scores. Reliability can also be affected by what Kelly Burling, Pearson's director of educator effectiveness, calls "cognitive load." That's the ability of the evaluator to keep several different categories in the rubric in mind when doing an observation. She recommends having administrators focus on a single trait that a district finds problematic, so they can refine their ability to look for evidence in a single domain..

**Click here to learn more about inter-rater reliability training from Pearson's Kelly Burling**

**Click here to hear Kelly Burling's recommendations about how administrators remain focused on scoring single traits**

Training—and retraining—are crucial to inter-rater reliability. Evaluators begin with extensive training in the instructional framework chosen by the district and learn how to discern a score of "2" from a 3 on the myriad indicators within the framework. They then become comfortable with the scoring process by initially rating teachers in video clips and build up to rating teachers in the classroom.

An evaluator's reliability is first assessed with the certification exam authorizing them to conduct official observations Evaluators are then required to recertify annually, and depending on their performance on the recertification exam, may be required to undergo further training to update their skills. Some districts have developed online video libraries of teaching practice, with clips that have been rated by master teachers, providing evaluators another source of assistance.

National Center for Teacher Effectiveness

Click here to learn how Kristan Van Hook of NIET recommends that districts monitor inter-rater reliability

The National Institute for Excellence in Teaching (NIET) recommends that evaluators work on teams within their districts, and keep an eye on average ratings. Their evaluation practice could include monthly team practice observations, in which evaluators could have a conversation about their collective understanding of what it means to score "2" or "4."

Some districts calibrate their scores on a quarterly basis to help ensure reliability.

Robert Ramsdell, vice president, Cambridge Education, recommends that evaluators do at least 10 practice observations before undertaking an official evaluation in the classroom.

To further calibrate the rating system, Cambridge Education is developing a "validation engine" on the data platform that receives observation data, to discern trends and check on the accuracy of raters. It also works in school districts and states to calibrate the rating scales, to make sure they are fair and equitable across several entities using the same instructional framework.

Click here to learn more about the validation engine from Cambridge Education's Robert Ramsdell

Tim Daly, president at The New Teacher Project (TNTP), warns against becoming too preoccupied with inter-rater reliability—having any two observers of the same teaching practice arrive at similar ratings. He says many differences in rating on particular indicators are inconsequential, when looked at in the totality of a teacher's evaluation. There may be differences on specific indicators, but a teacher's summative score, when taken from the average of all the ratings, would remain unchanged.

He equates variation in the scoring to the way umpires in baseball call balls and strikes. "You can adjust the strike zone, but you have to be within limits. If you are really off, you have a big problem. You should tolerate variation, and focus on the outliers."

Click here to hear Tim Daly of TNTP talk more about expected amounts of differentiation between scores

The human factor can play a role in rater reliability. There will be pressure on administrators to inflate scores, especially for those teachers with whom they have had longstanding relationships. Evaluators also need to guard against bias about certain classroom approaches involved the use—or non- use—of technology. "You must be aware of it, but don't be consumed by it," he said.