How teachers experience the evaluation process is related to how well-trained the evaluators are, the accuracy and consistency of their ratings, how effective the feedback is, and the level of support they receive for their improvement efforts. Many teachers look forward to being part of the process, appreciate and value the feedback they receive, and feel like their practice is continually improving. Overall, they describe their school's culture as being professional and respectful. What makes the difference?

In a recent TELL Survey (2015) Colorado teachers who plan to stay in their position see effective school leadership as having positive impact on evaluation compared to those leaving, and a majority of those staying believe the evaluation process improves instructional practice and is fair. The implication is that effective leaders support a climate of fair and consistent evaluations that help teacher improve their practice.

Fortunately, there are key elements within an evaluation system—widely accepted criteria, standards, and guidelines grounded in best practices—that are useful tools for creating and supporting positive cultures. When these elements are in place, agreement among evaluators is most likely to occur. And, when evaluation systems ensure there is consistency and fairness, the resulting positive culture of ongoing learning and teacher effectiveness can lead to increased student achievement.

What is Inter-rater Agreement?

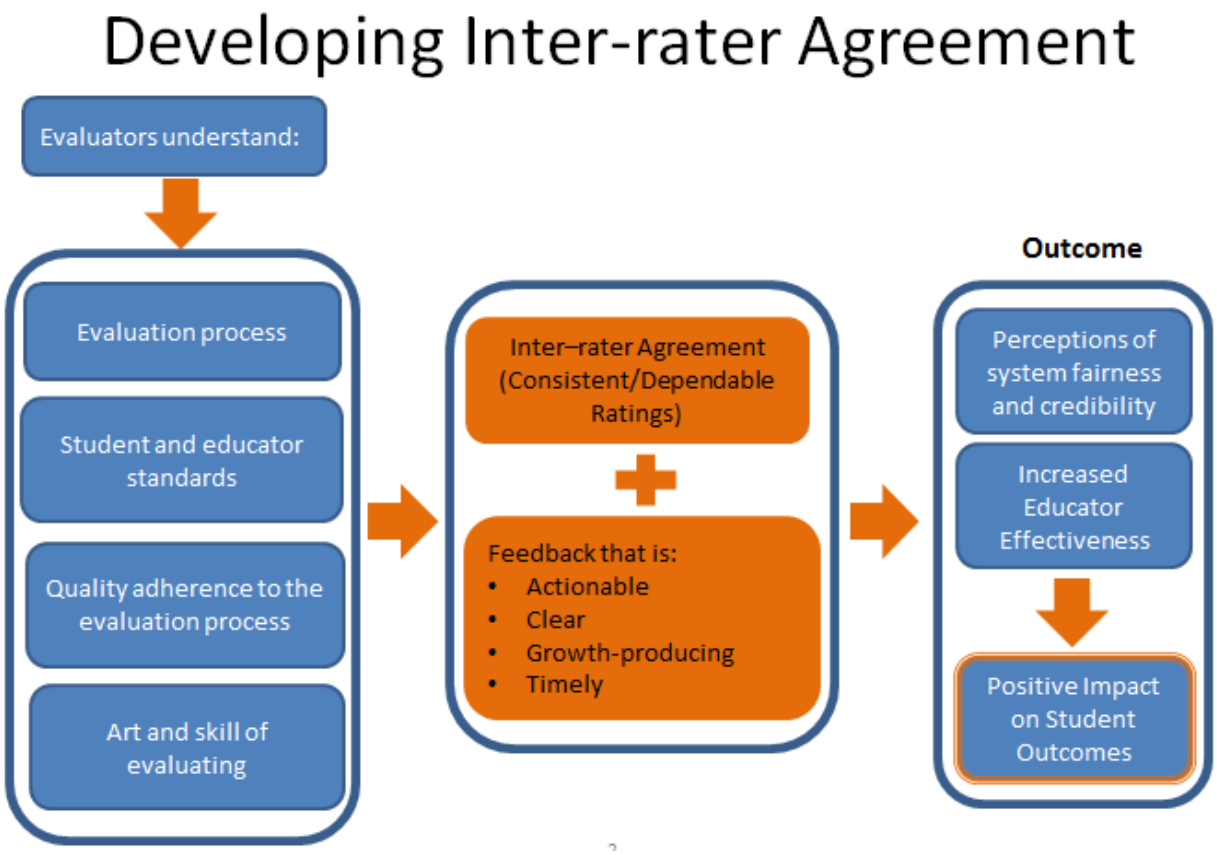
Because classroom observations can provide teachers with formative feedback to improve their practice, it is important to improve classroom observation techniques to ensure evaluators are consistently identifying high-quality teaching practice and identifying area for improvements. This critical work is where the topic of inter-rater agreement, or IRA, comes in. One way to understand IRA is to break down the jargon, beginning with the two terms you most often see in the research: inter-rater reliability and inter-rater agreement.

In education research, inter-rater reliability and inter-rater agreement have slightly different connotations but important differences. Inter-rater *reliability* is *t*he degree of agreement in the ratings that two or more observers assign to the same behavior or observation (McREL, 2004). In other words, when one rates a

teacher a three overall, the other consistently rates the same educator one point higher or lower. That means that the two raters have reliable scores, however they do not give the same score to the teacher. Inter-rater *agreement,* on the other hand, "measures how frequently two or more evaluators assign the exact same rating (e.g., if both give a rating of "4" they are in agreement)" (Graham, Milanowski, & Miller, 2012, p. 5.). Inter-rater agreement is the more useful term for discussing teacher evaluation because it tells us how frequently two or more evaluators assign an identical rating.

What comprises an effective IRA system?

The graphic below outlines each of the components of a strong IRA system with the ultimate goal of a positive impact on student achievement. We will explore each part of the system below.



Guiding Practices When Developing Inter-Rater Agreement Systems

Some techniques to use to ensure high levels of inter-rater agreement depend on the desired outcome. In the case of teacher evaluations, the ideal outcome is that multiple evaluators agree that a particular teacher's instruction on a given day meets the high expectations and rigor described in the state standards so that

students get the best instruction possible. It is this idea of agreement among evaluators that is so valuable to the process, and to achieve agreement, we turn to the research for guidance. There, we find that to maintain high levels of inter-rater agreement, the system must include ongoing professional development on the evaluation system.

Training, Training, and More Training

Recall the last time you decided to pursue professional development. The description of the training—just a day-long event—resonated with you. You arrived early that day to begin soaking in advice from the renowned speaker and at day's end, you were energized. But it didn't take long for most of the information from that invigorating day to wear off. There simply hadn't been enough time to practice, discuss, and learn a new process, and you haven't found time to embed your new learning into the day-to-day work.

Graham, et.al, (2012) recommends that training be embedded, linked to current practice, include follow-up and be delivered in a way that can be absorbed appropriately. Any technology used also needs to be of appropriate duration given the content to be learned. The best IRA systems will result from deep, sustained training around the evaluation rubric and the standards, elements, and professional practices and process. Specifically, training should include guidance from experts in how to use the rubric and practice on how to collect, record, to conducting evaluations (Kane, Taylor, Tyler, & Wooten, 2010). In addition, they should learn how to recognize bias that might have an impact on their evaluations. There should also be follow-up practice and calibration sessions and ways to embed the practice in an ongoing way.

The Evaluation Process

Guiding Practice: Provide training to all evaluators in the evaluation process. Ensure that evaluators understand the purpose and intent of each connection. Provide resources that support the implementation of the process so that evaluators can more readily implement each connection.

Let's take a look through the lens of State Model Evaluation System. Imagine that one evaluator follows all the connection points in the process and observes educators frequently to gather evidence and provide feedback, while another gathers only evidence from one formal evaluation, the scores generated from the evidence gathering might be significantly different.

Example: State Model Evaluation System Process:

Therefore it's imperative that the process is consistently applied among evaluators in a district and embodies the goals of evaluation; in this case, continuous improvement and increased student achievement, and those educators understand why each connection is important. A deep understanding ensures proper implementation. Developing resources to support implementation is helpful. One such resource is the Colorado State Model Performance Management System which provides a platform for collecting evidence and reporting with the State Model Evaluation System in mind.

Student and Educator Standards

Guiding Practice: Schedule enough time for evaluator training so participants can probe into the rubric (including the standards, elements, and professional practices). Include training on the expectations of the Colorado Academic Standards to become familiar with instructional strategies evaluators should expect to see. Use videos that have been master scored to support understanding. Build in time for participants to discuss the entire process.

Remember that training session where you were handed printed materials and told to follow along and ask questions as needed? For some, that approach is just fine, but for most learners, "seeing is believing." Videos allow us to see new requirements in action, and they teach what to do or what not to do. Videos not only make training more interactive, but when they are readily available, educators can access information on their own time and review it as often as needed or they can use a particular video to focus on just one aspect of the evaluation rubric.

Equally important, and to develop a shared understanding of what high-quality teaching looks like, evaluators must have opportunities to discuss observations with colleagues who use the same rubric and have access to a forum for sharing questions, insights, and tips with others (McClellan, Atkinson, & Danielson, n. d.). In addition, it is important for teachers to receive training so they understand the rubric and the process and can participate in providing their own feedback to evaluators. A report from the Reform Support Network (2013) describes the benefits of collecting teacher feedback: "1) giving teachers a sense of ownership of the process; 2) using the data to constantly improve the feedback that teachers receive; and 3) being able to tell a broader audience that teachers play a fair and meaningful role in the process" (p. 4).

Quality Adherence to the Evaluation Process

> **Guiding Practice:** Build systems to track steps in the process, monitor and analyze implementation and evaluation data, such the use of informal and formal data (e.g., how many times evaluators were in teachers' classrooms).

One way to ensure quality adherence to the evaluation process is to have periodic checks for understanding. During Administration team meetings, make evaluation part of the conversation. Look at district-wide data to find areas of concern. Have a dialogue around the data with the team to uncover possible root causes and find resources to help develop understanding. Highlight practices administrators are using that are helping them manage the evaluation process, like how and when they schedule and make time for observations.

As another potential way to improve agreement, research suggests that it is important to hold evaluators accountable for accurate ratings. One way to accomplish this is to review some of the evaluators' scores on artifacts or classroom observations, possibly by randomly double-scoring videotaped observations or artifacts. By comparing scores and talking about the evidence, evaluators can calibrate their thinking with each other. Using master-coded videos and artifacts helps ensure calibration beyond a school or district.

Art and Skill of Evaluating:

Use live practice often

> **Guiding Practice:** Do live evaluations with experienced evaluators and discuss what everyone saw, how the evidence and practice relate to the rubric, and what type of feedback might be given. Train and use independent observers, where possible, or make better use of staff who can serve in this role.

Live practice deepens understanding. It is similar to role-playing in that it is considered "experiential learning" and shares many of the same advantages, the first of which is to peak interest in the topic at hand. Jarvis, Odell, and Troiano (2002) point out that interest in a subject matter and understanding of the content both increase when experiential learning is introduced. They note that it encourages individuals to reflect on their own knowledge and understanding of a topic and that it requires participants to use the correct language, concepts, and arguments of the task, all of which deepen learning. For evaluators, this means confidently using the terminology found in the evaluation rubric and being able to provide feedback that teachers view as something they can act on.

Learn about evaluation bias

Guiding Practice: To help evaluators focus on the instructional practice(s) they are observing, include training sessions on the types of bias, their causes, and ways to guard against them.

Because observation, feedback, and evaluation inherently depend on human judgment, bias is an important issue to learn about. Learning why and when bias occurs is the first step in reducing it to ensure fair evaluations. For instance, research indicates that evaluators are more lenient when they know they will have to justify those ratings in a face-to-face meeting with the person being evaluated (Graham et al., 2012). At other times, the evaluator might have preconceived notions based on prior experience with the educator that interferes with the evidence from the teacher's practice. This might be considered to the teacher's advantage as in the Halo Effect or a disadvantage to the teacher. Either way, it's not truly in the best interest of the teacher or her students if the feedback is not based on an accurate assessment of the teacher's practice. There also are new concerns that underscore why it is imperative not to over or underestimate the quality of a teaching practice. A recent report from the Brown Center on Education Policy at Brookings (2014) states, "Our data confirm that such a bias does exist: teachers with students with higher incoming achievement levels receive classroom observation scores that are higher on average than those received by teachers whose incoming students are at lower achievement levels" (Whitehurst, Chingos, & Lindquist, 2014, p. 2–3).

Always follow up

Guiding Practice: Create opportunities for periodic re-calibration, assessing raters for "drift" and monitoring or re-training observers.

There is no substitute for follow up. After initial exposure to the rubric and the evaluation process, provide evaluators opportunities for ongoing learning, practice, reflection, and adjustment. According to Thompson and Goe (2009), it is in the follow-up phase of professional development where the new content has the potential to actually change teaching, learning, teachers, and schools. School and district leaders must work to provide opportunities for practice, reflection, and adjustment of observations and feedback. Videos are a particularly efficient means to reflect on teaching practice.

Why IRA Matters

Consistent and Dependable Ratings

Marketing, healthcare, and psychology are just a few disciplines, along with education, where inter-rater-agreement systems have long been used. As an educator, you might have been asked to be a scorer of essays written for an Advanced Placement test or a judge of a debate competition. In these situations, you quickly see the value of multiple raters who each assign the same scores to a single performance. Similarly, IRA systems for teacher evaluation can provide reliable, consistent data that support a shared vision of professional practices in the classroom. By using an IRA process, teachers and evaluators both can be confident that evaluation ratings are accurate and reflective of a teacher's success with students. This forms the basis of trust and meaning in an evaluation system.

Colorado educators should have confidence that the Colorado State Model Evaluation System, or any evaluation system, will result in fair, credible, and reliable ratings regardless of the person who conducts the evaluation. Not surprisingly, evaluation ratings for which IRA is strong are more credible and dependable than those for which it is lower. They are a better source of performance feedback and a reliable source of school- and person-level data for use in school or district decision making because they are more likely to reflect true strengths and weaknesses rather than a single evaluator's opinion of good educator practice.

Give teachers meaningful feedback

Guiding Practice: Make sure teachers are aware of the evaluation criteria ahead of time. Follow up with teachers after an observation and use the evaluation process as times to have a dialogue about evidence of teaching practices.

Following each observation, evaluators will be engaging in a conversation with a teacher they see almost every day. Learning to provide high-quality feedback in an open and honest way is one of the most powerful and exciting parts of the evaluation process. High quality feedback is consistent, clear, actionable and growth-producing. It has a great impact on the teacher when it's delivered in a timely fashion and meets these criteria. There are strategies evaluators can learn and practice to get better at providing meaningful feedback in a positive way while managing potential difficult conversations. Evaluators should keep in mind that most teachers feel that classroom observations can be a fair picture of their teaching (Goe et al., 2008), and most teachers want meaningful feedback.

A Closing Thought

There is mounting evidence that supports the connection between standards-based teacher evaluations and student achievement. Kane, Taylor, Tyler, and Wooten (2010) conducted studies in Cincinnati Public Schools to "test whether classroom observations—when done by trained professionals, external to the school, using an elaborated set of standards—can identify teaching practices most likely to raise achievement" (p. 2). They found that classroom observations and student achievement growth are related in substantial ways. Classroom observations have an undeniable upside--the potential of providing formative feedback to teachers that helps them improve their practice. Therefore, we must strengthen how classroom observations are measured.

References

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis.* Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved from http://files.eric.ed.gov/fulltext/ED521228.pdf

Graham, M. Milanowski, A., & Miller, J (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings.* Center for Educator Compensation and Reform. Retrieved from http://es.eric.ed.gov/fulltext/ED532068.pdf

Jarvis, L Odell, K. & Troiano, M. (2002, April). *Role-playing as a teaching strategy.* Strategies for application and presentation. Retrieved from http://imet.csus.edu/imet3/odell/portfolio/grartifacts/Lit%20review.pdf

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2010). *Identifying effective classroom practices using student achievement data* (NBER Working Paper No. 15803). Retrieved from http://www.nber.org/papers/w15803. (JEL No. I21, J45)

McClellan, C., Atchison, M., & Danielson, C. (n. d.) Teacher evaluator training & certification: Lessons learned from the measures of effective teaching project. *Practitioner Series for Teacher Evaluation*. San Francisco, CA: Teachscape.

Reform Support Network. (2013). *Promoting evaluation rating accuracy: Strategic options for states*. Washington, DC: U. S. Department of Education. Retrieved from http://www2.ed.gov/about/inits/ed/implementation-support-unit/tech-assist/evaluation-rating-accuracy.pdf

Thompson, M. & Goe, L. (2009). *Models for effective and scalable teacher professional development.* Princeton, New Jersey: Educational Testing Service.

Whitehurst, J., Chingos, M. M., & Lindquist, K. M. (2014). *Evaluating teachers with classroom observations—Lessons learned in four districts*. Washington, DC: Brown Center on Education Policy at Brookings.

Resources and Tools to Support Inter-rater Agreement

Administrators have long had many evaluation tools from which to choose, from informal *walkthroughs, to interviews and surveys, to scheduled formal classroom observations. Likewise, teachers have used self-reflective assessments, action research, or student growth data, among others.* Each has its advantages and disadvantages, and each can work well. The best kind of tools, of course, are the ones that help achieve the aim of ensuring a system that leads to increased teacher effectiveness. The following tools can assist you in putting the key elements in place in your district.

1. Colorado State Model Performance Management System

The Colorado State Model Performance Management System is an optional tool that will support districts in the implementation, data collection and effective use of the Colorado State Model Evaluation System. The performance management system includes electronic interfaces and data collection tools for the state model evaluation rubrics, measures of student learning/outcomes, final effectiveness ratings, and aggregate reports to support principals and district leaders to provide useful and actionable feedback and possible professional development opportunities for educators - See more at:
http://www.cde.state.co.us/educatoreffectiveness/copms#sthash.5tWUQ2Zz.dpuf

2. A Resource Guide for Deepening the Understanding of Teacher's Professional Practices

The *Resource Guide* can support observers and coaches in accurately identifying evidence of professional practices and can assist classroom teachers in reflecting on their teaching and planning for implementation of specific practices in their instruction. In addition, using the guide can lead to the development of a common language among Colorado school district employees as they analyze, reflect on, and plan instruction. Features of the guide include these:

- **Research** that provides a rationale for the professional practices in the Rubric and can be useful to anyone providing training on an element or providing feedback to a teacher.
- **Ideas**, **suggestions, and classroom examples** for implementation of a practice.
- **Planning/coaching questions** to support teachers' thoughtful decision making when planning instruction.
- **Resources,** such as articles, videos, websites, and supporting documents that provide further information about implementation of a practice.
- A **glossary**, **bibliography** for each Teacher Quality Standard, and an **alphabetical index** of all internal resource documents.