Washington State Teacher and Principal Evaluation Project

Maximizing Rater Agreement

Facilitator Guide

March 2015
Contents

Overview of the Training Module Series .......................................................... 1
  Audience .......................................................................................... 1
  Timing and Structure ...................................................................... 1
  List of Training Modules .................................................................. 2

Preparing for the Conducting High-Quality Observations and Maximizing Rater Agreement Modules .......................................................... 4
  Module Overview ........................................................................... 4
  Context ......................................................................................... 4
  Intended Outcomes ...................................................................... 5
  Agenda .......................................................................................... 5
  Equipment and Materials ............................................................... 6
  Prepare ahead of time: ................................................................. 6

Facilitator Guide .................................................................................. 7
  Welcome ......................................................................................... 7
  Connecting ................................................................................... 10
    Connecting Content .................................................................. 11
  Learning I: Rater Agreement Explained ........................................... 12
  Learning II: Rater Agreement Pitfalls ............................................ 15
  Learning III: Professional Learning to Maximize Rater Agreement .......................................................... 20
  Learning IV: Evidence-Based Feedback ........................................ 25

Implementing Segment ........................................................................ 31
  Implementing Activity II: Monitoring and Maintaining Rater Agreement .......................................................... 32
  Implementing Debrief ..................................................................... 33

Reflecting ............................................................................................ 33

Wrap-Up .............................................................................................. 34
Overview of the Training Module Series

This series of seven training modules is designed to prepare school leadership teams for the implementation of new educator evaluation systems through the following intended outcomes:

- Making the components of educator evaluation concrete and actionable for educators and their evaluators.
- Supporting district and school leadership teams in developing a common understanding of the new educator evaluation legislation, criteria, and frameworks, and the opportunities for professional growth and development.
- Provide participants with implementation tips and strategies to help schools make educator evaluation meaningful and doable.

The training module series will accomplish these goals through the use of consistent, standardized training materials, detailed facilitator guides, and participant handouts.

Audience

The audience for each module is district-level leadership teams of four to six people. Districts may bring some or all school leadership teams together for a single training session, or training sessions may consist of teams from several districts within a region.

The Office of the Superintendent of Public Instruction (OSPI) recommends that each district leadership team include the district administrators, principals, and two to three current classroom teachers representing a variety of subject areas, grade levels, and/or student needs. The modules are designed so that school leadership teams can facilitate abbreviated or complete versions of each module to other school staff (see the Timing and Structure section for more details). For districts that are interested in bringing these trainings back to their school sites, team members should be comfortable presenting information to a group of adult learners.

Timing and Structure

Each training module is three to five hours in length and includes interactive learning activities (some activities are optional, depending on the length of time allotted). Suggested homework assignments described at the conclusion of each module are intended to help participants extend and apply their learning and are designed to take about an hour.

The modules are organized into a four-part structure to help facilitators and participants pace the content appropriately. The four segments of each module are as follows:

- **Connecting**—Builds community, prepares the team for learning, and links to prior knowledge, other modules, and current work
- **Learning**—Describes key concepts and highlights various implementation scenarios; supports teams in applying knowledge and sharing ideas
- **Implementing**—Supports teams in problem-solving and planning next steps for schools and districts
- **Reflecting**—Engages participants in providing feedback, reflecting on learning, and closing the session

**List of Training Modules**

**Introduction to Educator Evaluation in Washington.** The basics of educator evaluation reforms, the evaluation criteria for teachers and principals, the four-level rating system, the state and local decision matrix, and a preview of the remaining modules.

**Using Instructional and Leadership Frameworks in Educator Evaluation.** An orientation to the components of instructional and leadership frameworks, how they are different from previous evaluation tools, and how they support identification of practice across a continuum. This module will provide a “jumpstart” into the three instructional and two leadership frameworks.

**Preparing and Applying Formative Multiple Measures of Performance: An Introduction to Self-Assessment, Goal Setting, and Criterion Scoring.** An overview of the types of measures required and supported by RCW 28A.405.100 and WAC 392-191A, the differences between measures and evidence, how to move beyond an observation-only evaluation system, and the benefits to the types of measures used in educator evaluation. This module includes criterion scoring guidance that has been informed by the instructional and leadership framework authors.

**Including Student Growth in Educator Evaluation.** An overview of goal setting for student growth, selecting classroom-based, school-based, district-based, and state-based tools, and using student learning data in educator evaluation. This module will offer a process for establishing student growth goals, examples of student growth goals, and a process for determining the change in student achievement between two points in time.

**Conducting High-Quality Observations and Maximizing Rater Agreement.** An overview of high-quality observation practices, with special emphasis on collecting evidence, strategies for maximizing rater agreement, and strategies for districts and school administrators to learn about and plan for maximizing rater agreement. This module will also address the important connection between rater agreement and providing high-quality feedback, in preparation for the Providing High-Quality Feedback for Continuous Professional Growth module (see below).

**Providing High-Quality Feedback for Continuous Professional Growth and Development.** An overview of examples of, and protocols for, how to provide feedback to teachers and leaders so that they continue to grow and improve in their practice; how
to engage faculty in these conversations; and strategies for connecting professional
development planning with evaluation outcomes.

**Combining Multiple Measures Into a Summative Rating.** While an evaluation rating
is often viewed as a measure of a single point in time, it is actually the culmination of a
comprehensive process of self-assessment, goal-setting, plan implementation,
dialogue, and reflection that unfolds over months. This module provides an
understanding of how to assess practice using multiple types of evidence and
performance rubrics in a thoughtful, comprehensive, reliable manner and follow the
process from OSPI to create an overall performance rating.
Preparing for the **Conducting High-Quality Observations** and **Maximizing Rater Agreement** Modules

**Module Overview**

This module is divided into two separate but interdependent modules. The first module, Conducting High-Quality Observations, focuses on the following:

- Providing participants the opportunity to examine expectations and promising practices for gathering evidence through classroom observation.
- Examining the role of observation in providing formative feedback and maximizing rater agreement.
- Reviewing the minimum observations requirements established for different categories of educators.
- Reviewing common sources of bias in observation and learning best practices for collecting observation data.

The second module, Maximizing Rater Agreement, provides participants the opportunity to:

- Examine expectations and promising practices for maximizing rater agreement for educator evaluation in Washington.
- Examine the OSPI’s working definition for rater agreement and practice strategies for avoiding common rating errors.
- Begin developing implementation plans for observation and rater training in their own districts.

Both modules are intended only to serve as an introduction and orientation to these topics. Districts should seek more in-depth training from other sources.

**Context**

Participants will have varying levels of knowledge about strategies and practices for conducting high-quality observation and achieving rater agreement in their district. Some may serve in districts that have devoted significant time to observer training and calibration using previous instructional and leadership frameworks; other participants may have relied on a “checklist” approach to classroom observation that was only loosely connected with a framework. Districts will also have differing levels of readiness for implementing observation and rater training plans for staff.

The content of this module is framed to accommodate a differentiated approach to audience knowledge. This module specifies activities that require the facilitator to know the level of audience knowledge and readiness. The facilitator will need to preselect
activities appropriate for the audience and make appropriate adjustments to the PowerPoint and materials based on those decisions.

**Intended Outcomes**

At the end of this session, participants will:

- Be able to describe the OSPI working definition of rater agreement and the stages for development.
- Be able to identify common rater pitfalls and strategies to overcome them.
- Understand the elements of high-quality training required to achieve maximum rater agreement.
- Explore strategies that support productive feedback conversations.
- Develop a district plan to maximize rater agreement.

**Agenda**

1. **Entry Task and Welcome (5 minutes)**
   a. Entry Task: Confidence Conversation (3 minutes)
   b. Welcome, Introductions, Overview, and Norms (2 minutes)

2. **Connecting (25 minutes)**
   a. Connecting Content (5 minutes)
   b. Connecting Activity: Snowball Common Terms (10 minutes)*
   c. Connecting Wrap-Up/Debrief (10 minutes)

3. **Learning (2 hours 35 minutes)**
   a. Learning Content 1: Rater Agreement Explained (10 minutes)
   b. Learning Content 2: Rater Agreement Pitfalls (15 minutes)
   c. Learning Activity 1: Strategies to Avoid Pitfalls (15 minutes)
   d. Learning 2: Wrap-Up/Debrief (10 minutes)*
   e. Learning 3: Professional Learning to Maximize Rater Agreement (10 minutes)
   f. Learning Activity 2: Example Rater Agreement Activity (75 minutes)
   g. Learning 3: Wrap-Up/Reflection (5 minutes)
   h. Learning 4: Evidence-Based Feedback (15 minutes)

4. **Implementing: Rater Agreement in Your District (1 hour)*
   a. Implementing Content 1: Planning for Implementation (15 minutes)
   b. Implementing Activity 1: Implementation Planning (40 minutes)
   c. Implementing Wrap-Up/Debrief (5 minutes)
5. Reflecting (5 minutes)
   a. Revisiting the Outcomes
   b. What’s Next/Homework

*Note: If you are attempting to complete all content and activities in a single 3-hour session, these activities are optional. The Implementing section can be assigned as “homework” (see slide 51). If you are planning for a single 5-hour session, you will want to build in breaks as necessary, and possibly extend Learning Activity 2 (see slide 36 for more information).

**Equipment and Materials**

**Equipment:** Laptop computer and projector

**Materials:**

Make a copy of the Participant Handout packet for each participant. Ensure you are using the packet that aligns with the district’s instructional framework.

Put the following materials on each table:

- Markers and highlighters (several per table)
- Standard size self-sticking notes (several pads per table)

Bring the following materials for use by you:

- Chart-paper markers (at least one of each color: red, blue, green, black)
- Chart paper

**Prepare ahead of time:**

- For slide 2: Entry activity charts (see directions for slide 2).
- For slide 11: Sets of index cards (one set per table) with the following terms written on them: rater agreement, consistency, accuracy.
- For slide 12: Two charts—one with definition of consistency and one with definition of accuracy.
Facilitator Guide

Welcome

Slide 2 is the entry task slide.

During this slide, remind participants to complete the instructions on the slide (teams talk and then a designee posts a color-aligned sticky dot for each statement on both charts).

Chart 1: Content

- **All** staff understand the shift between the State Criteria and previous evaluation.
- **All** staff frequently refer to the State Criteria and Instructional Framework when planning for instruction and/or school improvement.
- **All** staff are actively involved in aligning professional learning to State Criteria and Instructional Framework
- Staff have attended Stage 1 and Stage 2 training provided by an IFS or Framework author.

Chart 2: Culture

- A clear and shared vision for supporting evaluators is in place.
- There is a common and shared definition for rater agreement in our district.
- Teachers and administrators are comfortable discussing agreement and accuracy of ratings.
- The district provides structures (collaboration time) and resources for educators to share, learn, implement, and sustain rater agreement activities.

Entry Task: District Self-Assessment

- **Discuss:** Using the statements posted on the charts around the room, discuss your district’s level of practice.
  - Green: This is fully implemented in our district; we could help others with this.
  - Blue: This is partially implemented in our district; we have individuals who champion this work.
  - Yellow: We have begun working toward implementation; we need a lot more guidance on how to make this happen.
  - Pink: This has not been addressed in our district/department.
- **Identify:** Select one person from your district team to use the sticky dots to identify your district’s level on the posted charts.
Welcome participants, introduce yourself, and ask participants to briefly do the same with their first name and role. If the group is large, ask for a quick identification of district teams by district.

Explain the logistics for today’s sessions (restrooms, food/beverages, etc.).

This slide lists the agenda for the session.

**Explain:** As in previous modules, the agenda for this session is divided into four main sections. We will begin with some connecting activities to help build community, prepare everyone for learning, and link to prior knowledge, other modules, and your current work. We will then transition to the learning segment, where the key concepts for this module are explained including why rater agreement is important in providing feedback. Then you will have an opportunity to begin planning as a team for implementing a plan for achieving rater agreement in your district.

We’ll end with reflection time so that you can provide feedback and reflect on your learning, and then we’ll close the session.

**Explain:** The outcomes for today’s session are as follows: [read slide]. Each segment of the session is aligned to one or more of these outcomes, and the divider “transition” slide between segments articulates those outcomes.

As a reminder:

- **This module provides an orientation to the basic concepts.**

- **This module does not go into great depth about evidence relating to any of the specific instructional or leadership frameworks and instead leaves it up to the districts to seek additional training.**

- **This module assumes you have completed the modules that came before it and that you have a solid knowledge of the instructional framework and conducting observations.**
**Explain:** This module is part of a series of modules that OSPI has packaged together to support statewide implementation of new educator evaluation systems. Section 5 of Senate Bill 5895 states that “the Office of the Superintendent of Public Instruction must develop and make available a professional development program to support the implementation of the evaluation systems required by RCW 28A.405.100. The program components may be organized into professional development modules for principals, administrators, and teachers. These modules are intended to support the evaluation system model components as outlined here on this table from the TPEP website. These components are intended to fit together to create and support a system of professional growth for educators in our state.”

**Explain:** As a reminder, the six core principles for TPEP are as follows:

- Quality teaching and leading is critically important.
- Teaching and leading is work done by a core team of professionals.
- Teacher and principal evaluation models should coexist within the complex relationship between district systems and negotiations.
- Professional learning is a key component of an effective evaluation system.
- Evaluation systems should reflect and address the career continuum.
- An evaluation system should consider and balance “inputs or acts” with “outputs or results.”
Explain: Let’s have a brief review of the norms for today’s sessions. These were the norms we discussed during previous modules. [Read the norms on the slide and reference the chart paper posted on the wall with the norms listed.] You also have a copy of these in your participant packet to refer to throughout the session.

Any comments about these norms? Any to add?

Now, with these norms in mind, let’s get started with the module content.

Connecting

Slide 8 is the transition slide to the Connecting segment of the module.

Remind the group of the purposes of the Connecting segment (listed on the slide).

Explain: These modules aren’t intended to be the only training that you receive; they are meant to build awareness and a common foundation across the state. This module will give an orientation and basic understanding but you will need to seek and/or develop additional training specific to the instructional or leadership framework you are implementing in your district.

The content will not provide evaluators with the more comprehensive training needed for calibrating across observations to fully establish rater agreement among a group of evaluators. Additional, more extensive, training related to these evaluator skills is highly encouraged.
Connecting Content

**Explain:** We know that the new law requires that evaluators of both teachers and principals “must engage in professional development designed to implement the revised systems and maximize rater agreement.” We also know the Teacher and Principal Evaluation Project (TPEP) has relied heavily on the growing body of research, the framework authors, and practical input from practitioners in pilot sites to create a “working definition” of rater agreement.

To get us all on the same page and using common language, let’s take some time to think through common definitions for terms used in rater agreement work.

**Explain:** Turn to Handout 1 in your handout packet. The research and practical application of implementing this law is of primary concern for the Office of Superintendent of Public Instruction (OSPI), TPEP Steering Committee organizations, and the instructional and leadership framework authors.

In partnership with the three instructional framework authors and the Association of Washington School Principals (AWSP), OSPI will use the following working definition and steps: [See slide for steps.]

We are now at Stage 3 and today’s session will explore and unpack rater agreement.

**Explain:** Maximizing rater agreement is an important step in ensuring that new Washington educator evaluation systems will truly support professional growth and learning.

Rater agreement is a linchpin in the effectiveness of your new evaluation system. Research-based frameworks and excellent data collection practices will have little value if you cannot demonstrate rater agreement.
To start, we want to establish shared definitions so we’re all talking about rater agreement in the state the same way.

On your tables, you have three index cards, each with a different term written on it. In your table groups, define each term, and prepare to share your definitions in 5 minutes.

Allow 5 minutes to define terms and then explain the next steps of the activity. Give groups 5 minutes for each step and then process through a 2–3 minute whole group share out.

Post the charts and definitions on a wall.

**Explain:** According to OSPI, here are the definitions for each of these words. Can someone read the definition of rater agreement to the group? (Select a volunteer to read.) How does this definition align with your group’s definition?

Allow participants to share a few comments.

Can someone read the definition for consistency? (Select another volunteer for consistency and then one to read the definition for accuracy.)

Any final comments before we start to unpack these terms?

### Learning I: Rater Agreement Explained

<table>
<thead>
<tr>
<th>Key Terms</th>
<th>Consistency</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater Agreement</td>
<td>A measure of observer bias, indicating the extent to which observers assigned to the same observations of practice by another typical observer.</td>
<td>The extent to which the ratings follow the rows and columns of the agreement matrix.</td>
</tr>
<tr>
<td>Consistency</td>
<td>A measure of observer bias, indicating the extent to which observers assigned to the same observations of practice by another typical observer.</td>
<td>The extent to which the ratings follow the rows and columns of the agreement matrix.</td>
</tr>
<tr>
<td>Accuracy</td>
<td>A measure of observer bias, indicating the extent to which observers assigned to the same observations of practice by another typical observer.</td>
<td>The extent to which the ratings follow the rows and columns of the agreement matrix.</td>
</tr>
</tbody>
</table>
Explain: By demonstrating that your observers can achieve rater agreement, you will help ensure that:

- Educators can trust the new evaluation system.
- Educators will receive relevant, useful information for professional growth.
- The new system is legally defensible for personnel decisions. Nothing is guaranteed legally, but these practices are better than past practice and continually getting better. Making the effort to continually improve will go a long way."

Without demonstrating rater agreement, your new system will struggle in each of these areas.

Explain: Let’s look at what this definition looks like in practice. This table demonstrates two types of rater agreement: exact and adjacent. For component 1, two observers review the same classroom video and both rate the teacher’s practice as a 4, which matches the Master Score.

For component 2, Rater A scores the video as a 3 and Rater B scores it as a 2. Rater B is one point off from the Master Score and Rater A. This is an example of adjacent rater agreement.

Ask: Let’s look at Component 3. Is this adjacent agreement? Why or why not? Take 2–3 minutes to discuss with your table.

[Give participants 2 minutes to discuss and then continue.]

ANSWER: No, this represents “No agreement” because Rater A scored the component more than 1 point lower than both Rater B and the Master Scorer. This component demonstrates neither accuracy nor consistency across raters.

Ask: Let’s look at Component 4. Is this exact agreement? Take 2–3 minutes to discuss with your table.
[Give participants 2 minutes to discuss and then continue.]

ANSWER: No, although both raters scored the component as a 3 (i.e., they were consistent), the Master Scorer rated the component as a 1; therefore, the raters were inaccurate.

Explain: According to the OSPI definition, rater agreement refers to both the consistency and accuracy of the scores.

Explain: Rater agreement can be assessed at different levels (subcomponent/element level or the component/standard level). However, it is best assessed at the subcomponent or element level.

As the table on the slide illustrates, two raters can be more than 1 point different (beyond adjacent agreement) from the Master Score in four out of five subcomponents and still end up with an identical component score. The averaging process can mask significant rater disagreement. While this may not affect the teacher’s summative score, the feedback provided to the teacher based on the subcomponent/element scores would be misleading and inaccurate. For this reason, it’s important to assess rater agreement at the subcomponent/element level whenever possible.

Explain: Let’s take 5 minutes to discuss the following two questions with an elbow partner.

- What systems are in place for your district to focus on accuracy?
- Are these systems dependent on each other, or could a district start with agreement and then work towards accuracy?

Ask for a few volunteers to share the systems they identified in their conversations.
Learning II: Rater Agreement Pitfalls

Remind participants about the purposes of the learning segment and the specific outcomes aligned with this part of the session. The second learning segment focuses on common sources of rater agreement pitfalls and provides participants an opportunity to identify ways to identify and correct for errors, biases, and drift.

**Explain:** The research supports rater agreement as a necessary and critical component of the fidelity of educator evaluation systems. Strategies aligned with research-based activities can build observer skills and understanding and support deeper confidence in the overall observation process.

As you know, rating through observations requires evaluators to do a lot of things well. They must:

- **Be objective:** Record evidence that is free of “bias, opinion, and subjectivity” (Bell et al., 2013)
- **Understand the framework:** Thoroughly and deeply understand each component and indicator on the district framework instrument (Bell et al., 2013)
- **Gather and sort:** Collect sufficient evidence of practice as it happens in the classroom or school (McClellan, Atkinson, & Danielson, 2012)

**Explain:** Now that you have a basic understanding of what rater agreement and reliability are, let’s take a few minutes to consider what things affect rater agreement and reliability in teacher and principal performance. There are rater errors, rater biases, and rater drift.
Explain: There are different types of errors:

- Central tendency error
- Contrast effect error
- Halo error
- Potential error
- Leniency and severity errors
- Incident bias
- Recency bias
- Similar-to-me bias
- Rater drift

Take out Handout 2 to take notes as we discuss each of these in detail.

There are two terms on each slide in this section, and the definition of each term is provided.

Explain: Central tendency error occurs when a rater evaluates the observation using points on the middle of the scale and avoids extremely high or low ratings. Particularly during the implementation stage for a new system, raters are likely to feel pressure to be cautious and may want to inflate low ratings for teachers. To avoid conflict, the rater may simply err toward the middle of the scale. In a minute, we’re going to identify ways in which you can avoid this and other errors.

Contrast effect error occurs when a rater directly compares the performance of one educator to that of another educator. Your observers should be particularly vigilant about this error when they have multiple teachers on a focused evaluation all being evaluated on the same criteria. Because of the narrower focus of the observation, direct comparisons between teachers are more likely to occur.
Halo errors occur when a rater allows the ratings they select on one component/indicator to influence how they decide to rate another component or scale. For example, a rater may observe a teacher with excellent skills in delivering instruction. Because this is such a core aspect of a teacher’s practice, the rater may then “expect” the teacher’s performance on other components to be equally high, which may not be the case.

Potential error occurs when a rater gives higher or lower ratings to an educator than is warranted by the observation evidence because he or she believes the educator has (or does not have) the potential to be an excellent educator. For example, an observer who has an initial impression that an educator is not improving over time may artificially lower the educator’s ratings because the observer is anticipating stagnated performance. Similarly, if the observer’s initial impression of the educator is that he or she is dynamic and a fast learner, the observer may artificially inflate the ratings to meet this expectation for growth.

Leniency and severity errors occur when a rater gives mostly high (lenient) or low (severe) ratings to an educator in a manner inconsistent with the observation data collected. These are more common when raters lack a deep and thorough understanding of the framework (or parts of it) and attempt to rate without clear and explicit reference to evidence.

**Explain:** Recency bias can influence ratings because of how the observer recalls evidence. Raters tend to remember more recent events better than those that occurred in the past. For this reason, raters can end up weighting events that occur near the end of a lesson more heavily than those that occurred earlier in a lesson. This error reinforces the importance of taking high-quality running records of the lessons and completing ratings as quickly after the observation as possible.

“Similar-to-me” bias occurs when raters give higher ratings to educators who are similar to themselves.
and lower ratings to educators who are dissimilar. Teaching experience, content areas, race, cultural background—a wide range of factors could influence raters’ decisions. Raters will need to be thoughtful in considering how their life experiences may influence how they interpret the evidence collected.

**Explain:** Focusing on one or two incidents occurs when ratings are based only on a small sample of observation evidence that typically includes either very strong or very weak examples of practice. As a result, the educator receives an artificially high or low rating. For example, perhaps an observer was particularly impressed with how the teacher diffused a student’s misbehavior and rates them high on classroom management; however, the rater ignores contradictory evidence of student misbehavior the teacher ignored or did not address.

Rater drift happens naturally to even the most skilled observers. Rater drift is an inevitable outcome for observers without access to:

- ongoing calibration conversations involving real-life or video-based observation
- practice videos to refine skills

Rater drift is also much less likely when observers know that their ratings will be part of their professional growth conversation with their supervisor and part of a principal’s own evaluation.

**Explain:** To explore this a little more thoroughly, we’re going to do an activity. Take out Handout 3: Rater Agreement Pitfall Strategies. At your table, work with your group to match each common rating error with a possible strategy for helping evaluators to avoid the error. After you finish matching, discuss any other potential strategies that you have used or think might be effective.

You have 10 minutes to complete this activity.
Let’s take a look and see how we did. On this slide is the answer key. Score your handout. Did anyone get all of the answers correct? How many missed fewer than three? Would anyone like to share an additional strategy that came up at their table?”

Additional strategies to mention, if no one shares any:

- **A=** Contrast effect: Ask someone who did not collect the evidence to review it and determine if the rating aligns with the evidence and anchors.
- **B=** Leniency and severity errors: Do not try to be intentionally “easy” or “hard” in your ratings.
- **C=** Halo error: Remember that framework components are scored separately. Your ratings on one component should not influence ratings on another component.
- **D=** Potential error: Examine your bias toward or against a specific educator before observing them and control for the bias when rating.
- **E=** Rater drift: Conduct annual recalibration or recertification.
- **F=** Similar-to-me bias: Consider the observation evidence for each component separately and only use the information that is relevant to the component you are considering.
- **G=** Central tendency: Compare observation evidence with the behavioral anchors.
- **H=** Incident bias: Conduct multiple short observations in order to collect as much evidence as possible.
- **I=** Recency bias: Create ways to highlight or call out important evidence throughout the observation.

**Explain:** It is important to understand these pitfalls and consider them when collecting and analyzing evidence. This process is not scientific and will always include professional judgment and opinion, but to the extent we can minimize it, the better we are able to provide evidence-based feedback and support teacher improvement.
Learning III: Professional Learning to Maximize Rater Agreement

Remind participants about the purposes of the learning segment and the specific outcomes aligned with this part of the session. The third learning segment focuses on strategies and activities that can promote stronger rater agreement.

Explain: As mentioned earlier, the research and practical application of implementing this law is of primary concern for the Office of Superintendent of Public Instruction (OSPI), TPEP Steering Committee organizations, and the instructional and leadership framework authors. We are now at Stage 3, and today’s session will explore and unpack rater agreement.

Explain: Once we reach Stage 3, we do not graduate from Stage 3. We must revisit and review key learnings from Stage 1 and Stage 2 to avoid drift. Rater agreement is not ensured by a single training or certification test. Districts will continuously move evaluators to accuracy and consistency through ongoing calibration conversations involving real-life or video-based observation. Framework specialists do not provide training for this stage, although portions of Stage 2 training do have applicability toward rater agreement. Stage 3 should be integrated into the evaluation of principals (Criterion 5) and district administrators. Districts should be creating plans to provide for the ongoing professional development for all evaluators to demonstrate rater agreement.
**Explain:** Rater agreement is a characteristic of an evaluation system that requires ongoing calibration, opportunities to practice, and monitoring of ratings in order to be maintained. Evaluators cannot continue to achieve rater agreement over successive observations and successive years without these critical features.

We’ve just addressed the first bullet. Now we’re going to talk about ways to make the activities frequent and built around a common protocol or structure.

**Explain:** High-quality observation relies heavily on the skill of the evaluator who is conducting the observation.

Maximizing rater agreement is an important step in ensuring that Washington educator evaluation systems will truly support professional growth and learning.

Rater agreement is a linchpin in the effectiveness of your new evaluation system. Research-based frameworks and excellent data collection practices will have little value if you cannot demonstrate rater agreement. Under OSPI’s definition of rater agreement, the key point is that an educator’s observation score should be the same regardless of who is observing and rating the lesson.

When we think about rater agreement activities, we can keep in mind that there are many ways to design them. They can be several hours in length or they can be short, frequent activities. It’s important to note that frequency is the key to conducting the activities.
Explain: One crucial lesson from the original nine TPEP pilots was the importance of opportunities for informal calibration on an ongoing basis. Evaluators reported that these opportunities were some of the most meaningful and helpful in learning to use the framework in a more accurate and reliable manner.

Some pilot sites found that pre-existing administrator meetings, such as principal PLCs, were great opportunities to meet once a month to review and practice scoring videos on specific domains or components, to discuss scoring decisions, and to compare scoring with master scores.

Alternatively, opportunities for co-observation are useful calibration opportunities. Two observers watch the same classroom lesson, score the lesson separately, and then meet to compare and discuss scores. This approach is particularly helpful when you have a mixture of experienced and inexperienced evaluators.

Explain: Let’s take 5 minutes to think about how you might apply this lesson about the importance of ongoing calibration opportunities in your own district.

With others at your table, discuss what opportunities already exist for evaluators to meet and work together to calibrate their ratings. Try to identify at least two opportunities. You have 5 minutes to complete this activity.
Explain: Whatever format you select, it’s important to provide opportunities for observers to practice and receive feedback on how to gather observation data in a way that creates meaningful observation evidence (rather than opinion). To support accurate and consistent ratings, alignment of the evidence is critical. If evidence is not aligned properly, the ratings will not be accurate. Evaluators need reinforcement to ensure that in response to uncertainty the preferred strategy is to refer back to the scoring criteria. Last, observers need to be reminded and encouraged to use framework language while developing their rationale. A rationale can support identification of errors, biases, or drift, and can support conversations that are grounded in the framework language.

Explain: We’re going to walk through an activity that provides all of these components and a common structure in which to work through them. We will watch an 8-minute clip of a 4th grade math teacher. While the video is playing, collect evidence the way you normally collect evidence.

The video can be found here: http://www.insidemathematics.org/classroom-videos/public-lessons/4th-grade-number-operations-multiplication-division

Watch Introduction Part A and Problem 1 Part A.

Note to facilitator: You will need to make sure you have the correct handout and that page(s) from the corresponding framework are included in the handout packet.

Ask participants to turn to Handout 4: RA Activity Example (instruction framework). Participants will focus on the identified component on the handout only.

- Danielson framework: 3d. Using Assessment in Instruction
- CEL framework: A3. Formative Assessment Opportunities
- Marzano framework: 2.1. Interacting With New Knowledge

**Explain:** Individually, look through your notes and align evidence with the component.

Give participants 1-2 minutes to do this, then share out and chart and have participants jot down evidence in the first box on their handout.

Using this evidence, spend some time looking at the evidence and framework to individually determine a rating. Once you have identified a rating, complete the rationale box on your handout. This step is critical for a few reasons. First, it helps you learn the nuances between the levels of performance. It also provides you time to articulate the ratings so when you are having conversations with teachers you can reference this information and keep the conversation focused on the language of the framework and the evidence. Third, the more you are able to provide a rationale, the more likely it is that you are avoiding several of the errors and biases we discussed earlier.

Give participants 20 minutes to complete this portion. Circulate and attempt to make sure they are working individually.

**Explain:** Now that you have a rating and rationale, you are going to have a “norming” conversation in your table groups. Go around the table and share your rating and rationales. Identify any disagreements and talk through them as a table. Come to a consensus on a rating as a table group.

Give participants 15 minutes to complete.

Create a chart and tally up the ratings as they are shared. Ask each group to provide their rationales and build consensus in the room.
**Explain:** We talked about making sure rater agreement activities were regularly scheduled and even embedded in professional learning activities. This activity took about an hour, but let’s spend a few minutes talking about how you could conduct this activity in 30 minutes and how you might be able to conduct the activity in 90 minutes or more.

Suggestions could include:

30 minutes: Have evidence already typed up and aligned and spend 30 minutes rating and developing rationales.

90 minutes or more: Focus on more than one component or spend time after the rationale is determined to role-play the feedback conversation with the teacher.

**Learning IV: Evidence-Based Feedback**

Remind participants of the other module on feedback. The next section will provide insight on why rater agreement is critical in having evidence-based conversations about practice.

**Explain:** So once you have the ratings right, there’s a need to provide feedback. Research suggests that observation-based feedback has the potential to make a powerful impact on student learning.

Regardless of whether the training relies on video or live observation, a critical feature of high-quality training is the provision of immediate feedback and continued coaching through discussion of ratings.

Observers need to be able to explain and understand the ratings, and not merely be able to record an accurate rating.
Without this level of understanding, evaluators will struggle to provide the teacher with useful, actionable feedback during post-conference conversations.

Let’s consider a couple of questions: When you think about postobservation conferences in your district, how confident are you that:

1. Your teachers are taking an active role in the conversations?
2. Principals are preparing teachers to participate and take ownership over the process?

You can either ask for volunteers to respond or ask people to do a thumbs up, sideways, or down.

Consistency and accuracy in your ratings can support the conversation in many ways. Being sure you collected evidence and aligned correctly will support deeper conversations about practice.

The bulk of the feedback should be focused on evidence collected during the observation. This helps avoid several big dangers in postobservation conferences:

1. Loose interpretation: Evidence-based feedback is the key factor that differentiates an actual observation conversation from a conversation over different opinions about instruction.
2. Subjectivity: Using evidence effectively during the conversations keeps things squarely centered on a common, shared understanding of effective instruction, and depersonalizes the feedback process.
3. Emotion: Evidence-based feedback can also help remove some of the emotion from the evaluation process.
Explain: When providing evidence-based feedback, evaluators need to learn to incorporate the language and vocabulary of the framework into their feedback. This helps build and reinforce a shared understanding of good instruction and also ensures that the framework remains the objective point of reference in the conversation.

Explain: The framework on the slide is adapted from information provided in the Chicago study. As you see here, low-level questions require little response from the teacher. Medium-level questions require some teacher response but tend to focus on completion of tasks and requirements. High-level questions, on the other hand, prompt back-and-forth conversations between the teacher and evaluator and really focus on instructional practice. These questions prompt teachers to reflect on their practice and can lead to much more constructive conversations. With a partner, generate three additional examples of high-level questions that evaluators might ask during a feedback conversation.

A study of teacher evaluation implementation in Chicago found that principals generally dominated the conversation, speaking 75 percent of the time in postobservation conferences (Sartain, Stoelinga, & Brown, 2011).

The same study also found that only 10 percent of questions asked by evaluators during postobservation conferences were high level and promoted discussions about instruction (Sartain, Stoelinga, & Brown, 2011).

My Teaching Partner study: This program provided focused, observation-based instructional feedback twice per month and produced student achievement gains of 9 percentile points (randomized controlled study) (Allen et al., 2011).
Cincinnati Study: This longitudinal study found that student performance improved the year a mid-career teacher was evaluated and even more in subsequent years (controlled for experience, type of students) (Taylor & Tyler, 2012).

To wrap up any conversation, it’s important to have some actionable next steps so there is a clear focus and goal.

Provide participants an opportunity to share their examples with the whole group.

Having ratings ahead of time helps you prioritize what to ask in the conference and how to structure questions to promote deep thinking and action.

**Explain:** As Hill and Grossman note, many evaluation instruments are broad rather than context-specific. During feedback conversations, evaluators may need to translate the text of the framework or other evaluation measures into what the teacher might do in his or her own classroom. From there, evaluators and teachers should develop a list of next steps for improvement, such as implementing a new strategy in the classroom or seeking out targeted professional development. Whenever possible, a follow-up plan should be determined so that the teacher not only implements the feedback in the classroom but also receives additional coaching and feedback on the area of need, either from the evaluator or from another support professional, such as an instructional coach. If possible, the evaluator should help the teacher practice or model the practice during the postobservation conference, or if this is not feasible, point the teacher to another colleague who can model the practice, or to print and online resources.
Explain: In order to have honest conversations, it’s important to have a culture of trust that embraces feedback. Take a few minutes and discuss at your table how you’ve built the culture, how rater agreement can help cultivate it, and what job-embedded strategies are in place to develop and support those skills in your district.

In your table teams, discuss:

- What strategies have you found successful in building a culture that embraces feedback and reflective conversations about teaching and learning?
- How does rater agreement help build a culture of reflection and continuous improvement?
- What job-embedded strategies are in place for evaluators to practice specific skills?

Ask each group to share one strategy they discussed.

Explain: There are additional statewide resources available to you, which we’re going to discuss next. Who is familiar with eVAL? Who uses it frequently?

eVAL is easy to navigate, simple to use, and supports any locally adopted framework for teacher and principal evaluation. eVAL has a number of roles, but the primary use is to support dialog between supervisor and supervisee, reflection about an individual’s practice, and the collection of evidence of effective practice. eVAL serves as a central storage and reporting tool for many aspects of the evaluation process. The primary users of eVAL are teachers and principals. A summary of their uses of eVAL is summarized below.

- Key features to mention:
  - Easy access to the video library for eVAL users within the eVAL tool
  - A search tool that helps users locate videos by grade level, subject area, strategy, and average rating
- Custom playlists for each user that can be used for practice session with others
- The ability to invite others to join in a practice session, document an observation, and use the framework
- The use of the familiar Notes Editor that is the same as the real Notes Editor used for authentic observations
- A method to review the practice sessions with others, allowing them to analyze the video and compare their perspectives
- The power of the eVAL library: … reaches beyond the practice session. It promotes dialogue between principals and teachers, allowing them to share their notes and scores, and to compare their perspectives on the practice observation session.

*Videos provided by Baker Evaluation Research Consulting ([http://www.bercgroup.com](http://www.bercgroup.com))

The benefit to a resource library like this is that it supports the activities that move evaluators toward the 3rd stage of rater agreement. [Reference slide 10]. Think about the rater agreement activity we completed to observe 4th grade math [reference slide 37/handout 4]. This kind of activity is one that could be used repeatedly with different groups of evaluators with different grade level and content focused videos from a resource like eVAL.

**Explain:** If you’d like more information on these, please contact your ESD for training on this eVal resource.

<table>
<thead>
<tr>
<th>Resource: eVAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Anyone can view videos or protocols at:</td>
</tr>
<tr>
<td>- Users can use a generic log in to practice rater agreement in the eVAL Sandbox:</td>
</tr>
</tbody>
</table>

Slide 48
Explain: The video exercise you just completed can be a helpful way to start a discussion with evaluators on what effective feedback looks like. For additional resources on providing instructional feedback, we recommend the resources included here.

- **Learning-Focused Supervision: Developing Professional Expertise in Standards-Driven Systems** (Lipton & Wellman, 2013)
- **Principal Evaluator’s Toolkit for the Instructional Feedback Observation** (American Institutes for Research, 2012)
- **Leveraging Leadership: A Practical Guide to Building Exceptional Schools** (Bambrick-Santoyo, 2012)
- **The Art of Coaching: Effective Strategies for School Coaching** (Aguilar, 2013)

Explain: There are also other modules that you might want to revisit or schedule if you haven’t already participated in them:

- **Introduction to Educator Evaluation in Washington**
- **Using Instructional and Leadership Frameworks in Educator Evaluation**
- **Conducting High-Quality Observations and Maximizing Rater Agreement (Part 1)**
- **Providing High-Quality Feedback for Continuous Professional Growth and Development**

Implementing Segment

Slide 51 is the transition slide between the Learning and Implementing segments. The Implementing segment provides time for districts to begin the planning process for developing a plan for ongoing rater training, calibration, and monitoring.

Facilitator note: If you are shortening this module, the next section could be assigned as “homework” to district teams.
Implementing Activity II: Monitoring and Maintaining Rater Agreement

**Explain:** You will have the next hour to begin developing a plan for monitoring and maintaining rater agreement over the long term in your district.

Begin by taking approximately 20 minutes to read “High Fidelity: Investing in Evaluator Training” (Handout 7). This brief will provide you with additional information and examples to help inform your planning process and a self-assessment that will help you determine where to focus. It might also be helpful to review the “Rater Agreement in Washington State’s Evaluation System” (Handout 1).

**Explain:** In just a minute, your team will have some time to plan, but first, we want to share a few other things that Washington districts are doing around rater agreement.

The North Mason School District partnered with a Danielson expert to create a video protocol, similar to the protocol we used during this session.

**Explain:** In Central Kitsap, former Superintendent Hazel Bowmen developed “Instructional Reviews Protocol.”
**Explain:** Use the remainder of your time to engage your team in conversation about the Implementation Tool (see, Handout 6: Implementation Planning) and to begin developing your district’s own plan for monitoring and maintaining rater agreement.

I will check in with everyone in about 20 minutes and facilitate moving along to the next task as necessary to help everyone manage their time. During the debriefing I will ask teams to share a decision they made (key, preliminary, a change, etc.) and an immediate next step. We can all help each other be a little more accountable for actions.

**Implementing Debrief**

**Explain:** Teams have worked hard in this session to actively engage in productive tasks. Let’s have each team share two things to debrief our implementing tasks: 1) One decision you made today (could be a key decision, a preliminary decision, etc.) and 2) one of the immediate next steps you are taking when you return to your district.

Each team shares out.

**Reflecting**

Slide 57 provides the transition to the Reflecting segment of the session.

The Reflecting segment is intended to be a time for participants to share feedback about the session and plan for the next session.
To determine how well we’ve met our outcomes for today, turn to a partner and:

- Describe the OSPI working definition of rater agreement and the stages
- Share one rater pitfall and the strategy to overcome it.
- Identify a best practice or strategy that could be used to maximize rater agreement.
- Explain a strategy that would help administrators develop skills in providing feedback.
- Discuss how your district plan includes elements listed above.

**Explain:** Some possible ideas for how you can extend the learning you engaged in today to those from your district who cannot be here:

- **District or school teams:** Use your observation notes to practice scoring additional indicators in your framework and discuss ratings to achieve agreement.
- **Identify specific components or dimensions you think will be particularly hard for your observers to score.** Prioritize those components or dimensions in your ongoing calibration and practice sessions.

**Wrap-Up**

Thank the participants for coming and make sure you have inserted your email address and other contact information on this slide so they can get in touch with any follow-up questions.