MET project

POLICY AND PRACTICE BRIEF

Building Trust in Observations

A Blueprint for Improving Systems To Support Great Teaching

Use dialogue
Include internal thinking
I wonder
I thought to myse
Add a simile: compare using like or as
screamed, hollered, raced, peered
Show details and description by adding adjectives or adverbs
ferocious amazing lovely
quickly

BILL & MELINDA GATES foundation

ABOUT THIS DOCUMENT: States and districts can use this tool to create their own plans for continual improvement of an observation system, no matter where they are in their implementation. Included are action steps to improve observation rubrics, observer training, observer assessment, and monitoring. A planning process is described to assess current status, determine next steps, and plan for sustained improvement. The MET project will release more detailed guidance on how to implement the action steps in a series of briefs beginning in summer 2014.

ABOUT THE MET PROJECT: The MET project is a research partnership of academics, teachers, and education organizations committed to investigating better ways to identify and develop effective teaching. Funding comes from the Bill & Melinda Gates Foundation. For more information, and other MET project resources, see: www.metproject.org.

ABOUT THE AUTHORS

- **Jess Wood,** a former middle school teacher, led development of an online observer training system for the District of Columbia Public Schools (DCPS). Now a policy advisor at EducationCounsel—an education law, policy, strategy, and advocacy organization—she works with states, districts, and education organizations on issues related to teacher evaluation and professional development.
- **Cynthia M. Tocci,** a former middle school teacher, is an executive director in research at ETS. She was the senior content lead for Danielson's Framework for Teaching during the MET video scoring project and for the Teachscape Focus observer training and certification system.
- **Jilliam N. Joe** is an associate research scientist in the Teaching, Learning, and Cognitive Sciences group at ETS. She provided psychometric and design support for the MET video scoring project and currently provides research and implementation direction for the Teachscape Focus system.
- **Steven L. Holtzman** is a research data analyst in the Data Analysis and Research Technologies group at ETS. He has served as a research lead on the MET video scoring project and several other MET studies.
- **Steve Cantrell,** a former middle school teacher, is the chief research officer responsible for research and evaluation on K–12 education investments at the Bill and Melinda Gates Foundation, where he has co-directed the MET project.
- **Jeff Archer** is a communications consultant specializing in school improvement issues. He has led the MET project's efforts to communicate the implications of its research to practitioners.

ACKNOWLEDGEMENTS: Four practitioners with extensive experience in implementing high-quality classroom observations provided invaluable advice on the development of this resource: Colleen Callahan (Rhode Island Federation of Teachers and Health Professionals), Sara Heyburn (Tennessee Department of Education), David Steele (retired from Hillsborough County Public Schools, FL), and Jonathan Stewart (Partnerships to Uplift Communities [PUC] Schools). Consultant Mark Toner and KSA-Plus Communications provided editorial and design assistance. Others at DCPS made significant contributions to the content of this document, especially Stephanie Aberger, Michelle Hudacsko, Kim Levengood, and Scott Thompson. Tim Matheney of the New Jersey Department of Education provided helpful feedback on the blueprint's utility.

TELL US WHAT YOU THINK: This resource will be updated and augmented based on reader input. Email questions and suggestions to info@metproject.org. Include "Trustworthy Observations" in the subject line.

JUNE 2014

Contents

INTRODUCTION
Build Capacities that Build Trust

Classroom observations hold great potential to improve teaching and learning. In an evaluation and feedback system based on multiple measures, observations can clarify expectations for teaching, support teachers in elevating their practice, and provide essential information for key personnel and professional development decisions.

Moreover, when teachers receive regular, actionable feedback on their practice—rather than being left alone to assess their own progress—they are better able to make the instructional shifts called for by new college and career readiness standards, such as the Common Core State Standards.

No matter how far along a state or district is in the implementation of classroom observations, this blueprint can help plan for continued improvement.

But these benefits can easily be undermined by poor implementation. School systems that fail to take the necessary steps to ensure consistency in a climate of support run the risk of producing bad information that leads to distrust and bad decisions. When observers give conflicting feedback to teachers, it damages the credibility of evaluation and provides nothing with which to inform support for improved instruction.

Fortunately, what needs to be done is not a mystery. Experts and early adopters have learned a great deal in recent years about how to build trust in classroom observations. This document distills those findings into a series of action steps to develop and enhance each part of an observation system needed to produce trustworthy results. Together, these steps form a blueprint for improving observation systems to support great teaching.

No matter how far along a state or district is in the implementation of classroom observations, this blueprint can help plan for continued improvement. To those who are in the early stages, the action steps will suggest how to build on lessons learned from a pilot while planning for sustained improvement in the years ahead. To those who are further along, the same action steps will point out areas for refinement and areas for reinforcement to ensure the soundness of current structures. States may use the document to guide local efforts, improve state models, and prioritize areas in which to build district capacity.

TRUSTWORTHY OBSERVATIONS ARE:

- **Consistent.** Results vary little by observer or lesson.
- **Unbiased.** Results don't reflect personal or pedagogical preferences.
- **Authentic.** Expectations are clear and reflect best practice for effective teaching.
- **Reasonable.** Performance standards are challenging but attainable.
- **Beneficial.** Teachers get actionable feedback and support for success.

Key Components of a Trustworthy Observation System

Often when people talk about observations they refer only to the tools and procedures used by evaluators. But what makes observations trustworthy is a set of system components that work together to support evaluators in using those tools and procedures correctly. Trustworthy observations are the result of a proven observation rubric, carefully scaffolded observer training, assessment of observer accuracy, and ongoing monitoring of observations (see **Figure 1**).

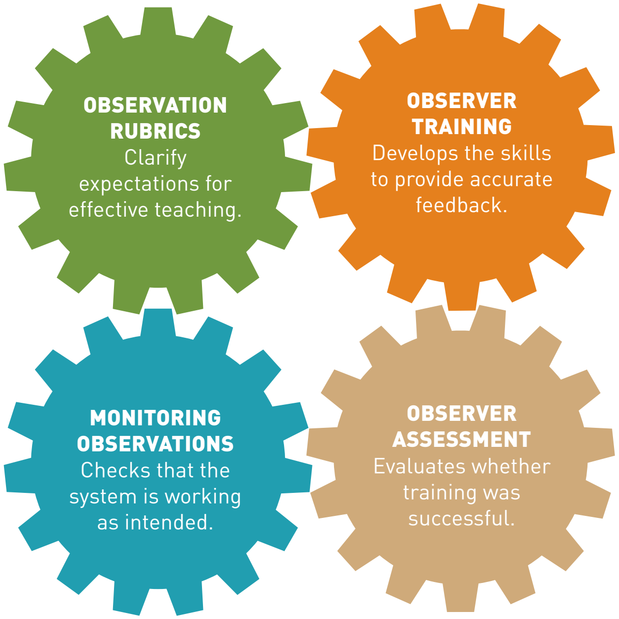
It may not be initially apparent that trustworthy observations depend on these components. Past instructional experience might seem sufficient to ensure that evaluators consistently identify effective teaching. A simple way to test this is to ask a few evaluators who haven't been trained to independently score a lesson video using their district's observation rubric, and then compare the scores they gave and the reasons that they gave them. Chances are that the ensuing discussion will reveal significant disagreements about the level of performance demonstrated, the meaning of the rubric, and even what behaviors were observed.

In fact, some districts use this exercise with observers to make the case for training. The point is not to question the value of experience but to expose the need to develop consistency, without which accurate feedback and fair evaluation is not possible.

While the components in Figure 1 support observer agreement that builds trust, they also support continual improvement of the system. An observation rubric serves as the basis for observer training, but the training of observers also reveals parts of the rubric that need clarification. Observer assessment builds confidence that observers have mastered requisite skills, but it also exposes the need for enhancements in training. Monitoring observations provides information to assess and improve all parts of the system, including a district's efforts to support more effective teaching.

Often when people talk of observations they refer only to the tools and procedures used by evaluators. But what makes observations trustworthy are a set of system components that work together to support evaluators in using those tools and procedures correctly.

Figure 1. A Trustworthy Observation System


OBSERVATION RUBRICS
Clarify expectations for effective teaching.
OBSERVER TRAINING
Develops the skills to provide accurate feedback.
MONITORING OBSERVATIONS
Checks that the system is working as intended.
OBSERVER ASSESSMENT
Evaluates whether training was successful.

These components don't emerge fully formed. Each requires the development of specific knowledge, tools, and processes. To be sure, states and districts can shortcut their efforts by adopting or customizing existing tools. Indeed, doing so when possible makes sense given the resources needed to implement tools even after they're developed. But no matter how much or how little is created from scratch, implementation entails a significant learning curve.

How to Use this Blueprint to Plan Continual Improvement

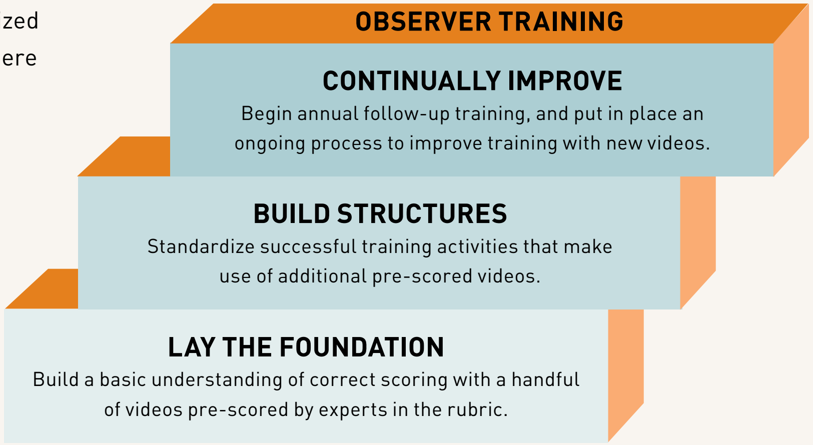
The action steps in this blueprint are sequenced to build the capacities that support a trustworthy observation system (see **Figure 2**). They begin with a set of foundational steps to forge the basic understandings needed to create the structures for quality implementation. Plans are then put into place to consider refinements and updates on an ongoing basis. Improvement at each step is informed by feedback and data.

Starting with a solid foundation saves the need for extensive rebuilding. Where a structure already exists, the foundation may need shoring up. As suggested by the examples in Figure 2, a field test of how to train observers with pre-scored video supports the soundness of subsequent training programs. But an existing training program might need shoring up with foundational steps to build a better understanding of the pre-scoring process to ensure that observers are normed to the right standard.

Starting with a solid foundation saves the need for extensive rebuilding. Where a structure already exists, the foundation may need shoring up.

Figure 2. Building Capacity

The action steps in this blueprint are organized in three phases of construction, as shown here in this summary for observer training.



OBSERVER TRAINING

CONTINUALLY IMPROVE

Begin annual follow-up training, and put in place an ongoing process to improve training with new videos.

BUILD STRUCTURES

Standardize successful training activities that make use of additional pre-scored videos.

LAY THE FOUNDATION

Build a basic understanding of correct scoring with a handful of videos pre-scored by experts in the rubric.

To help states and districts create their own improvement plans, the sequence of action steps in this blueprint is organized within 16 essential activities (see next page). Research and the experience of early adopters suggest that these activities are what drive the components of a trustworthy observation system. To give an example, feedback from teachers and observers supports rubric clarity. To give another, multiple opportunities to practice scoring make for effective training. A continual improvement plan should address each of the 16 activities.

A process to create such a plan is outlined on the bottom of the next page. To get a clear picture of a system's current status it's important to compare the work thus far to the action steps in the blueprint for each activity. This should begin with a review of the foundational steps, no matter how far a state or district is in implementation. Addressing foundational steps not yet taken should be a first priority. After that, other unaddressed action steps should be accomplished in order.

For those just starting to implement observations, the result will be a plan that first lays the foundation across all activities in what amounts to a pilot of the whole system. Most states and districts, however, will find that they have addressed more of the action steps for some activities than others and that in some cases they need to go back and address foundational steps where structures already exist.

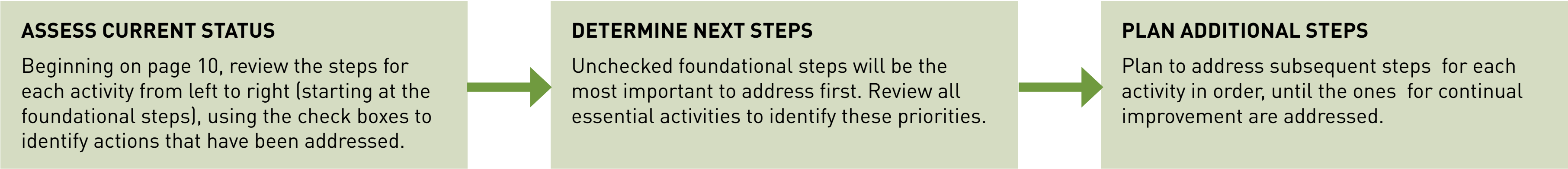
Addressing all steps in the blueprint should result in sustained improvement. In places where that's the case, a commonly understood language about teaching has taken hold. Teachers and evaluators see the criteria for effective teaching as clear, reasonable, and sound. They see the payoffs in terms of better practice and better decisionmaking. In short, they see a trustworthy system that supports great teaching.

ESSENTIAL ACTIVITIES IN A TRUSTWORTHY OBSERVATION SYSTEM

Research and the experience of early adopters suggest that the quality of an observation system depends on these essential activities. The pages cited clarify a sequence of action steps to build the capacity to address each, from a set of foundational steps to steps for continual improvement.

Observation Rubrics	
p. 10	**Aligning expectations.** Build buy-in for a set of commonly understood indicators of effective teaching.
p. 11	**Ensuring applicability.** Limit indicators to behaviors that can be expected in all lessons and that observers can reasonably track.
p. 11	**Ensuring clarity.** Leverage language and structure to support easy comprehension of each indicator.
p. 12	**Evaluating validity.** Check for evidence that results discern among teachers based on how well they promote student learning.
p. 13	**Soliciting feedback.** Ask teachers and observers how well the rubric supports consistency and instructional improvement.
Observer Training	
p. 16	**Pre-scoring video.** Use the rubric to determine benchmark scores and score rationales for videos of teaching.
p. 17	**Explaining rubric.** Provide an overview of the rubric's basis, structure, key features, and terms.
p. 17	**Minimizing bias.** Make observers aware of their biases and of ways to counter the possible effects of those biases on scoring.
p. 18	**Supporting practice.** Develop accuracy through explicit instruction, modeling, and practice scoring.
p. 19	**Modeling feedback.** Illustrate how to give teachers productive feedback based on observations.
Observer Assessment	
p. 22	**Determining tasks.** Create a scoring activity that mirrors what observers will do in the classroom.
p. 22	**Defining accuracy.** Set a minimum standard for scoring proficiency.
p. 23	**Establishing consequences.** Clarify what happens when observers fail to demonstrate sufficient accuracy.
Monitoring Observations	
p. 26	**Verifying process.** Inspect to see if observation procedures are followed.
p. 26	**Checking agreement.** Make sure observers maintain their accuracy.
p. 27	**Evaluating support.** Assess efforts to improve instruction.

HOW TO CREATE A PLAN FOR CONTINUAL IMPROVEMENT



ASSESS CURRENT STATUS

Beginning on page 10, review the steps for each activity from left to right (starting at the foundational steps), using the check boxes to identify actions that have been addressed.

DETERMINE NEXT STEPS

Unchecked foundational steps will be the most important to address first. Review all essential activities to identify these priorities.

PLAN ADDITIONAL STEPS

Plan to address subsequent steps for each activity in order, until the ones for continual improvement are addressed.

Clarify Expectations for Effective Teaching

At the heart of a trustworthy observation system is a well-designed rubric. A rubric outlines a common language for instructional practice that gets teachers, instructional coaches, administrators, and central office staff on the same page about what good teaching looks like. This has a profound effect on teachers' practice. What's in a rubric will shape what teachers do in the classroom.

A rubric also plays a central role in building the credibility of a new observation system. It is the basis of consistency and accuracy in scoring. While effective implementation also is critical for establishing legitimacy, the text of a rubric will be a teacher's first exposure to a new system. When teachers read the rubric, they need to be convinced that it's a fair set of expectations that they can agree with.

What's in a rubric will shape what teachers do in the classroom.

For all these reasons, an observation system built on a poorly designed rubric will fail in many respects. No amount of training will produce consistent scores if the rubric is unclear as to how to distinguish among different aspects of teaching and different performance levels. Nor will feedback lead to better student outcomes if the rubric emphasizes teaching behaviors that are unrelated to student learning. A well-designed rubric is the result of research, careful construction, and continuous improvement.

But even though it's critical to start with a sound rubric, it's just as important to understand that rubrics necessarily evolve. Their use—in pre-scoring videos for observer training and in evaluation—will suggest refinements that enhance consistency and validity. Moreover, as expectations for student learning change, so will the expectations for teaching that a rubric should emphasize.

The action steps on the following pages build from the initial determination of a rubric to an ongoing process of gathering information to identify aspects of a rubric for possible improvement. Reviewing these action steps may reveal foundational steps yet to be addressed for existing tools. For example, a rubric might be in use but teachers and evaluators haven't had the chance to weigh in on the clarity of its key terms. A plan to improve an observation system should prioritize any unaddressed foundational steps to accomplish first. Subsequent steps should be accomplished in the order shown.

KEY TERMS

Observation rubric. An observation instrument that outlines the criteria for different levels of teaching performance.

Rubric component. The level of rubric organization at which scores are assigned (e.g., use of questioning, checking for student understanding). Sometimes referred to as a *dimension* or *element*.

Rubric indicators. Descriptions of specific behaviors indicating different aspects of each component (e.g., wait time and cognitive demand for use of questioning).

Validity. The extent to which evidence supports a particular use of evaluation results as appropriate, such as discerning among teachers based on how well they promote student learning.

Embedded throughout the guidance in this blueprint is the caution to avoid non-essential changes to a rubric. Even small adjustments may have unintended consequences on teacher behavior or on scoring consistency. Proven rubrics are generally best left as-is until their use reveals problems.

What the blueprint doesn't detail are the steps to develop an entirely new rubric. The knowledge required to do so could fill an entire book. Any state or district that attempts to build a rubric from scratch will need to tap significant technical expertise and should allow at least a year for initial development. Given time and resource limitations, the most viable option for those just starting to implement observations will be to start with an existing rubric with evidence to support its validity.

ESSENTIAL ACTIVITIES FOR IMPLEMENTING OBSERVATION RUBRICS

Action steps to address each activity are on the following pages.

Aligning expectations. Build buy-in for a set of commonly understood indicators of effective teaching.

Ensuring applicability. Limit indicators to behaviors that can be expected in all lessons and that observers can reasonably track.

Ensuring clarity. Leverage language and structure to support easy comprehension of each indicator.

Evaluating validity. Check for evidence that results discern among teachers based on how well they promote student learning.

Soliciting feedback. Ask teachers and observers how well the rubric supports consistency and instructional improvement.

An Example of ENSURING CLARITY

The DCPS observation rubric reflects several features aimed at clarifying the distinctions among different aspects of teaching and different performance levels. The tool, the Teaching and Learning Framework, incudes nine components (called "TEACH Standards"), each scored based on two to five discrete indicators. Below are the indicators for scoring a teacher's demonstrated ability to check for student understanding.

CHECK FOR STUDENT UNDERSTANDING

	HIGHLY EFFECTIVE	EFFECTIVE	MINIMALLY EFFECTIVE	INEFFECTIVE
KEY MOMENTS	The teacher checks for understanding of content at all key moments.	The teacher checks for understanding of content at almost all key moments.	The teacher checks for understanding of content at some key moments.	The teacher checks for understanding of content at a few or no key moments.
ACCURATE PULSE	The teacher always gets an accurate "pulse" at key moments by using one or more checks that gather information about the depth of understanding for a range of students, when appropriate.	The teacher almost always gets an accurate "pulse" at key moments by using one or more checks that gather information about the depth of understanding for a range of students, when appropriate.	The teacher sometimes gets an accurate "pulse" at key moments by using one or more checks that gather information about the depth of understanding for a range of students, when appropriate.	The teacher rarely or never gets an accurate "pulse" at key moments by using one or more checks that gather information about the depth of understanding for a range of students, when appropriate.

Discrete indicators capture distinctly different aspects of the same component of teaching. Here, one relates to how often a teacher checks for student understanding when doing so would provide important information, and another relates to how often a check for information actually yields information with which to better address students' needs. Use of such discrete indicators also is meant to support more specific feedback to teachers.

Consistent scaling and language. Descriptions of performance levels for each indicator vary only interms of frequency, duration, or quality. Rubric guidelines specify that "All"=100%, "Almost All"=80–99%, etc.

ACTION STEPS TO IMPLEMENT CLASSROOM OBSERVATION RUBRICS

ALIGNING EXPECTATIONS. *Build buy-in for a set of commonly understood indicators of effective teaching.*

LAY THE FOUNDATION

□ **Identify a research-based rubric that aligns with state teaching and learning standards and with a vision of good practice articulated by teachers and school leaders.**

It's easier to support observations when the underlying expectations reflect what you believe is effective instruction.

- Convene stakeholders to identify common beliefs about what constitutes effective teaching (e.g., modeling and guided practice, checks for understanding). Keeping this group small and guided by clear structures helps avoid an unwieldy process.
- Consider adopting an existing rubric with a research base that reflects the identified beliefs, recognizing that to develop a new rubric is a significant undertaking and that resources may be better spent on the implementation process. Only build a new rubric if sufficient time and technical capacity are available to ensure a psychometrically sound instrument.

BUILD STRUCTURES

□ **Provide opportunities for teachers to see video examples of teaching that align with rubric components and performance levels.**

Seeing examples demystifies the observation process and gives teachers a concrete picture of what's expected of them in the classroom; see the Observer Training section on p. 14 for more information about videos.

- Provide teachers with the same rubric review material and the same types of pre-scored videos and score rationales used in observer training so that teachers understand how scores are determined based on objective evidence. (For action steps to pre-score video, see p. 16.)

CONTINUALLY IMPROVE

□ **Communicate to teachers and observers the substance, rationale, and process used to determine any changes to a rubric.**

Lack of understanding leads to suspicion and confusion.

- Documents like FAQs should clarify the need for the change, how the decision was made, and how it will affect scoring (e.g., higher-order questions are given additional weight in scores for questioning technique to better align with new college and career readiness standards for student learning).

ACTION STEPS TO IMPLEMENT CLASSROOM OBSERVATION RUBRICS

ENSURING APPLICABILITY. *Limit indicators to behaviors that can be expected in all lessons and that observers can reasonably track.*

LAY THE FOUNDATION

□ **Ensure that the rubric only describes components of effective teaching that should be observable in all lessons and that represent a manageable number of criteria for which observers should collect evidence.**

Consistent scoring is difficult when observers are overtaxed with too many criteria and when they must decide whether a behavior should have been evident in a particular lesson.

- For observations, criteria should only be behaviors that can be seen or heard and that don't depend on the observer having additional knowledge of the teacher or students.
- Consider how many indicators, components, and performance levels observers can reasonably track at the same time. Experience has led some rubric developers to streamline their tools to include no more than 10 components.
- Review each criterion in a rubric and ask, "Are there situations in which we wouldn't expect to see this in a typical lesson?" This should be tested as part of a pilot.

BUILD STRUCTURES

□ **Make rubric revisions or provide guidance to address concerns raised after use in observation that some criteria are not always evident and/or that the rubric includes too many criteria to score.**

Applicability to typical situations won't be fully evident until observers use the rubric to score actual lessons.

- Consider eliminating criteria for which evidence was sometimes absent (e.g., you can't expect to score teachers on use of technology if using technology isn't appropriate for developing student understanding in every lesson).
- Consider collapsing multiple rubric components into one if they include nearly the same criteria (e.g., classroom management and maximizing use of instructional time).
- Districts lacking authority to revise a rubric should raise such concerns with the rubric developer while also providing guidance to observers on handling such situations.

CONTINUALLY IMPROVE

□ **Augment the rubric's criteria for teachers of special populations, like special education students and English language learners, and supplement rubric descriptions with examples of what certain criteria might look like in non-core subjects, like art and physical education.**

While the core elements of effective teaching apply to all classrooms, specialization requires added skills and the ability to adapt the fundamentals of good teaching for different kinds of student learning.

- Convene groups of highly respected teachers of special populations and non-core subjects to review the rubric and suggest augmentations.

ENSURING CLARITY. *Leverage language and structure to support easy comprehension of each indicator.*

LAY THE FOUNDATION

□ **Verify that rubric terms, distinctions, and structures are clear to teachers and observers.**

Clarity increases the chance that all observers will read and apply the rubric the same way.

- Avoid vague terms and quantities (e.g., "students are generally engaged") in favor of observable and quantifiable indicators (e.g., "most students respond to questions during the course of the lesson").
- Ensure similar criteria are scaled across performance levels for each scoring component (e.g., "there are significant periods of time when students are idle" and "there are brief periods of time when students are idle").

BUILD STRUCTURES

□ **Set policies for how to determine scores when faced with evidence of varied performance.**

Consistency requires that observers weigh evidence in the same ways when determining scores.

- Make clear when some pieces of evidence should outweigh others (e.g., give more weight to the effectiveness of a teacher's checks for understanding than to how often the teacher checks for understanding).

CONTINUALLY IMPROVE

□ **Establish a clear process to periodically and carefully consider changes to the rubric and/or how it is explained to observers and teachers.**

Small changes can affect scoring quality in unexpected ways, so resist the temptation to revise rubrics without compelling evidence of the need to do so.

- Field test any potential changes to explanations of rubric criteria with observers to make sure they support consistent scoring.

ACTION STEPS TO IMPLEMENT CLASSROOM OBSERVATION RUBRICS

EVALUATING VALIDITY. *Check for evidence that results discern among teachers based on how well they promote student learning.*

LAY THE FOUNDATION

□ **Collect data from rubric developers and early participants in observations to see if students learn more when taught by teachers with higher observation scores.**

An important piece of validity evidence is the relationship between teachers' observation scores and measures of their students' learning. Observations should promote teaching practices that contribute to student learning.

- When choosing a rubric, ask developers for results of validation studies on the correlation between teachers' scores and student learning measures, such as value-added using standardized tests. Typical correlations are often in the 0.2–0.3 range (where 0 means no correlation, and 1 means a perfect correlation).
- Conduct an early validation study to compare teachers' observation scores with measures of student learning in their classrooms based on standardized tests the state or district uses.
- If overall trends in observation scores are unrelated to student learning gains, it may signal the need for better observer training or that the rubric itself does not capture teaching that supports student learning as measured by the assessment.

BUILD STRUCTURES

□ **Begin annual collection and analysis of validity evidence to determine whether, across the state or district, teachers' observation scores continue to show a relationship to measures of their students' learning.**

The relationship between a teacher's observation scores and measures of their students' learning may strengthen as observers become more skilled in applying a rubric. But that relationship also could erode if teachers change their practice in ways that result in higher observation scores but do little to improve student learning in their classrooms.

- If the relationship to student learning weakens, determine whether the problem is particular to specific parts of the rubric. If so, the rubric component may need revision or observers may need new guidance on how to score it.
- Keep in mind, however, that a lack of validity evidence also may be the result of inadequate observer training.
- Any time a rubric is revised, training must be updated so that observers understand how to apply the new criteria.

CONTINUALLY IMPROVE

□ **Continue annual collection and analysis of validity evidence, validating observation scores against new measures of student learning as they are adopted.**

Observation scores produced using a particular rubric may show a stronger or weaker link to student learning when different assessments are used to measure student learning.

- A weaker link between observation scores and student learning may indicate the need to revise the rubric to emphasize certain teaching practices that will better support the kind of student learning required by the new assessments.
- Eventually, if validity becomes so weak as to warrant a new rubric, a new tool should be piloted and a whole new training system should be developed to align with the new expectations for effective teaching.

ACTION STEPS TO IMPLEMENT CLASSROOM OBSERVATION RUBRICS

SOLICITING FEEDBACK. *Ask teachers and observers how well the rubric supports consistency and instructional improvement.*

LAY THE FOUNDATION

□ **Use surveys, focus groups, and informal discussions with early participants in observations to identify problematic aspects of the rubric that may have unintended consequences.**

Users are the best source of information about whether a rubric is having the desired effect of promoting consistency and good teaching practice.

- Prior to adopting an existing rubric, ask current users in other school systems about their experience.
- Ideally as part of a pilot, identify any overly specific criteria (e.g., teachers felt they had to address a specific number of learning styles spelled out in the rubric, regardless of whether it was appropriate to the lesson).
- Also identify any confusing language and distinctions (e.g., observers were unclear about when a question counts as a check for understanding and when it counts as a use of questioning to build student understanding).

BUILD STRUCTURES

□ **Begin annual survey and listening sessions with teachers and observers across the system to gauge support for the rubric.**

Over time, teachers and observers should see the tool as increasingly clarifying, helpful, and sensible. A survey during this phase provides a benchmark.

- Ask to what extent the rubric reflects good teaching practice and clarifies a set of reasonable expectations.
- Ask to what extent the rubric supports useful feedback.
- Ask whether parts of the rubric remain unclear or seem to be problematic.

CONTINUALLY IMPROVE

□ **Report changes in support for the rubric based on annual survey and listening sessions, and report how implementation is changing based on that feedback.**

The best way to build support and to keep getting useful feedback is to show you take it seriously.

- Give serious consideration to changes likely to improve consistency and validity, and make it clear when the suggestions of teachers and observers result in a change in tools or procedures.
- If support does not increase, hold additional listening sessions to probe why.

Develop the Skills to Provide Accurate Feedback

Observation is highly challenging: A classroom is a complex and unpredictable environment, and a lesson may include thousands of interactions. But accurate feedback and fair evaluation demand that any observer focuses on the same small number of behaviors to reach the same conclusions that any other observer would draw, if scoring correctly. Compounding the challenge is the fact that observers come to the task with their own personal and professional biases, of which they may not be aware.

A well-designed rubric helps to mitigate this challenge by calling out a few key teaching components and a few criteria for each performance level. But without training, even the clearest rubric will be applied differently by different observers.

The only way to confidently train evaluators to assign the correct scores is with examples of teaching for which the correct scores have been determined. For this reason, a central feature of observer training is video of teaching that has been pre-scored by expert observers in a process called master coding, which also establishes the correct justification for each score. These videos are then used for illustration, practice, and assessment.

Without training, even the clearest rubric will be applied differently by different observers.

Given the role that pre-scored videos play in setting the gold standard for accuracy, it's critical that pre-scoring involves quality-control checks to ensure that the assigned scores are indeed correct, based on the rubric.

But observer training involves more than watching and scoring. Observers must hone a set of supporting skills, like taking and organizing notes efficiently. They need to know what kinds of behaviors might relate to each component of teaching. They need fluency in the fine art of providing feedback. And they need to internalize these skills to the point where they can apply them quickly and correctly in the moment.

KEY TERMS

Accuracy. The extent to which observers are able to assign the correct score to a lesson using a particular rubric.

Inter-rater agreement. The extent to which multiple observers assign the same score to the same lesson or teacher. Also called *rater-agreement.*

Reliability. The degree to which scores are free from influences such as who gave the scores, when they were given, and what was being taught to which group of students.

Bias. Internal factors unrelated to the quality of teaching practice that may influence an observer's scoring decisions—either positively or negatively.

Evidence collection. A process in which observers record behaviors in the classroom without interpretation, typically through note-taking.

Master coding. A process in which experts in an observation rubric (*master coders*) review video of teaching and determine the correct scores, and the correct rationales for those scores, so that the video then may be used to train and assess observers for accuracy. Also called *anchor rating* or *pre-scoring.*

Reconciliation. A process in which multiple master coders (often in pairs) agree on the correct scores and on the correct evidence to support those scores.

Of all the components in a trustworthy system, observer training may entail the most capacity building for a state or district. Building a library of pre-scored video takes time and expertise that must be developed. Only through iteration does training become effective. Even once quality training is in place it takes practice for observers to apply their new skills with fidelity. The action steps in the pages that follow build this expertise on a foundation of basic understandings about what it takes to train effectively.

A state or district that adopts existing rubrics may be able to incorporate training offered by the developer. But most will need to create at least some training on their own or with others. Regardless of who delivers training, those who implement the observation system should ensure it's built on the right foundations. But they also should expect observer training to evolve as it grows.

ESSENTIAL ACTIVITIES FOR BUILDING OBSERVER TRAINING

Action steps to address each activity are on the following pages.

Pre-scoring video. Use the rubric to determine benchmark scores and score rationales for videos of teaching.

Explaining rubric. Provide an overview of the rubric's basis, structure, key features, and terms.

Minimizing bias. Make observers aware of their biases and of ways to counter the possible effects of those biases on scoring.

Supporting practice. Develop accuracy through explicit instruction, modeling, and practice scoring.

Modeling feedback. Illustrate how to give teachers productive feedback based on observations.

An Example of SUPPORTING PRACTICE

These excerpts from training on how to score a teacher's **USE OF QUESTIONING** show some of the activities in the online training system created by ETS and Teachscape for observers who took part in the MET project's study of classroom observation.

TRAINING ACTIVITY	EXCERPT FROM TRAINING
Rubric review. For each component, training points out the language distinguishing each performance level and gives written examples.	**Critical Attributes of Level 3 (out of 4):** • Most questions have multiple possible answers. • Discussions enable students to talk to one another without ongoing mediation.
	Possible Examples: Teacher asks: "What are some things you think might contribute to . . ."; "Michael, can you comment on Mary's idea?" Michael responds to Mary.
Modeling. Trainees review short clips and are told what evidence aligns with each performance level for a teaching component.	**Video A.** This is a 3 because the teacher's questions create a discussion among students. (Teacher: "What do you know about segregation? What does the word 'apart' mean?") . . .
	Video B. This is a 4 because the students themselves ask high-quality questions of each other. (After one student reads a passage, others call out, "What do you mean by that?") . . .
Practice scoring. Trainees score short videos after which they learn the correct scores and their rationales.	Your Score: 3 Actual Score: 2
	Your score is too high. **This is a 2** – the teacher's questions require single answers. All discussion is between the teacher and students. One student asked a question, the teacher told him to wait and did not return to answer it. The teacher used generic prompts, "Any questions?" **This is not a 1** because the questions relate to the lesson objective.

ACTION STEPS TO BUILD OBSERVER TRAINING

PRE-SCORING VIDEO. *Use the rubric to determine benchmark scores and score rationales for videos of teaching.*

LAY THE FOUNDATION

□ **Pre-score a starter set of videos using at least one pair of expert observers (master coders) who provide score rationales grounded in the language of the rubric.**

The only way to confidently train observers to score correctly is with examples of teaching for which the correct scores are determined.

- To build credibility, consider recruiting an initial cohort of master coders from among instructional leaders who are highly respected and influential.
- When master coders struggle to agree on scores, determine if the rubric is unclear, if one coder is misinterpreting the rubric, or if the video should not be used.
- As a quality-control check, have experts in the rubric review score rationales for alignment with the rubric criteria. Does the evidence cited align with the rubric's criteria for the score given?
- Master coding begins when preparing to pilot a rubric and continues during the pilot to provide additional pre-scored video for subsequent implementation.

BUILD STRUCTURES

□ **Before using videos, have them scored by a second set of expert observers to make sure they assign the same scores.**

Quality controls ensure that problematic videos and codes are not used in training.

- This quality control check on scores could happen simultaneously with scoring by another pair of coders, or it could take place after.
- Make sure each expert observer continues to score independently before comparing notes with a partner. This process is called reconciliation and it often works best in pairs; with more than two coders there's a temptation to reach consensus rather than agreement on what scores are best supported by the evidence.

CONTINUALLY IMPROVE

□ **Establish an ongoing process for master coding new video and recruit master coders from among the most accurate observers.**

Training videos should be replaced when they become outdated or when better examples are available. Assessment videos need replacing after many observers have seen them.

- Assign someone the task of determining when specific videos should be replaced by reviewing the videos and getting feedback from observers. Reasons might include changes in instructional technologies or certain teacher or student behaviors that prove distracting.
- Consider expanding the videos used in training to include teachers of a variety of subject areas, grade levels, and students.

LAY THE FOUNDATION

□ **In a starter set of videos, include at least one short video example of each rubric component at the middle performance levels (levels 2 and 3 of a 4-level tool).**

When starting to build capacity, it makes sense to prioritize videos that help observers distinguish between middle performance levels—those are the levels at which most teachers perform but where the distinctions are less obvious.

- Make sure all teaching components are covered in the starter set. Observers may lack confidence if asked to score teaching for which they haven't seen any examples.
- Coding additional videos early on in the process of implementing observations may not be a good use of resources. If the rubric changes significantly after the initial use, then video will need to be recoded.
- Including a few examples of high-level performance will help raise observers' expectations (especially if the examples are drawn from the local context).

BUILD STRUCTURES

□ **Add video examples of other parts of the rubric, prioritizing ones that address the biggest challenges observers encountered in the first round of training.**

Examples that clarify what's proven to be most confusing will get more bang for the buck in terms of increased accuracy.

- Consider where rangefinder examples (a high 2 or low 3) may be especially important to call out what distinguishes between levels.
- Look for mini-segments (maybe one minute long) that illustrate terms that proved problematic for observers (e.g., the meaning of "accurate pulse," used to describe whether a check for understanding was effective).

CONTINUALLY IMPROVE

□ **Complete a video library with enough examples to allow for sufficient observer agreement on all parts of a rubric.**

By this phase, supports should be in place so that observers can recognize every component and performance level.

- Multiple examples of the same part of the rubric can help observers see how the same expectations can look different in different classrooms, grade levels, and subject areas (e.g., two teachers check for understanding effectively, but one uses exit cards and another a turn-and-talk).
- A bank of medium-length videos (approximately 15 minutes long) that include multiple components of the rubric can further build accuracy and confidence by giving observers additional opportunities to practice scoring multiple components simultaneously.

ACTION STEPS TO BUILD OBSERVER TRAINING

EXPLAINING RUBRIC. *Provide an overview of the rubric's basis, structure, key features, and terms.*

LAY THE FOUNDATION

□ **Go over the rubric's structure and key terms, and make a case for scoring consistently with the instrument.**

Evaluators not accustomed to using a rubric may feel their own knowledge and experience is sufficient to ensure quality scoring.

- Explain how inconsistency undermines efforts to improve teaching and learning.
- Summarize the tool's theory of instruction (e.g., "Students succeed when challenged and supported in a conducive climate").
- Point out the common threads and variations in criteria across performance levels for each component (e.g., "Teacher always allows enough wait time" and "Teacher never allows enough wait time"). Discuss how evidence for each level would look different within the same components.
- Define key instructional terms and quantities (e.g., What is a higher-order question? What is meant by "most" or "few"?).

BUILD STRUCTURES

□ **Improve the overview of the rubric based on feedback from early participants and any changes to the rubric.**

The extent to which training succeeds in developing understanding won't become clear until some observers are actually trained.

- Survey observers at the end of training on what parts of their experience were most and least helpful in understanding the rubric.
- If needed, revise training for new teacher and leader induction to incorporate enhanced rubric explanations.

CONTINUALLY IMPROVE

□ **Establish a process to annually consider improvements to the rubric overview training based on feedback from observers, teachers, and trainers.**

Changes to content should clarify the expectations in the rubric but not change those expectations (unless the rubric itself has changed).

- Field test any potential changes to rubric explanations with observers to make sure they are interpreted as intended.
- Avoid unneeded changes as too many will confuse observers.

MINIMIZING BIAS. *Make observers aware of their biases and of ways to counter the possible effects of those biases on scoring.*

LAY THE FOUNDATION

□ **Explain to observers what is meant by bias in the context of observations and why it's important to address.**

Everyone has personal and professional preferences that they are not fully aware of and that could lead to scoring that's inconsistent with a rubric.

- Make three points: everyone has biases; we can't eliminate them but we can reduce their impact on our scoring; and awareness of biases helps an observer to score accurately.
- Explain common bias factors (e.g., speech and instructional methods) and provide examples of each (e.g., disfavoring the vernacular and favoring lots of student talk regardless of quality).

BUILD STRUCTURES

□ **Train observers to identify and counter their own biases.**

The importance of countering biases increases as evaluation carries greater significance for teachers.

- Prompt observers to self-reflect by asking them to rate the importance of different instructional methods and to write down things they would see in a classroom that would cause them to think favorably or unfavorably about the lesson. Doing this individually and anonymously promotes honest self-reflection.
- Encourage observers to keep lists of their own biases so they know when they need to make sure their biases are not affecting scores.

CONTINUALLY IMPROVE

□ **Establish a process to annually consider improvements to bias awareness training based on feedback from observers and those who train and supervise them.**

Observers will tell you when strategies are working and when others are needed. A big-picture view can identify systemwide issues.

- Give observers opportunities for feedback on bias awareness training at the end of training and after they have observed teachers in the classroom.
- Solicit feedback from trainers and supervisors on any systemwide trends they see that may need to be addressed in training (if many observers struggle with particular biases).

ACTION STEPS TO BUILD OBSERVER TRAINING

SUPPORTING PRACTICE. *Develop accuracy through explicit instruction, modeling, and practice scoring.*

LAY THE FOUNDATION

□ **Using pre-scored video, have observers record evidence for each teaching component in the rubric and assign scores.**

Observers need repeated practice to apply a rubric correctly.

- Prior to this process, explain the rubric, its key terms, and the types of evidence that would align with each teaching component (e.g., "If a teacher claps her hands and students stop talking, that's evidence of classroom management").
- Allow observers to compare their scores and rationales with correct ones produced by master coders. Attending to the right evidence and the correct interpretation and judgment of that evidence is as important as assigning correct scores.
- Consider the training formats (e.g., group, individual, and online) that best support this practice, keeping in mind that a flexible approach is best when starting out.
- If existing training is not available, new training may be piloted by having a large group practice together with the same videos, with three or four observers scoring and reporting out on each criterion of a teaching component. Start with a clear-cut one (e.g., classroom management).
- If limited resources demand prioritizing, focus the first round of training on components that are most likely to be confused or to drive improvements in practice (e.g., most teachers know how to effectively manage classrooms, but many could ask more probing questions).

BUILD STRUCTURES

□ **Codify, enhance, and build out effective practice activities from initial use so all observers experience the same high-quality preparation.**

Consistency in training promotes consistency in practice.

- Organize training into manageable chunks that cover a few teaching components at a time, and for each one, build from the rubric criteria to evidence collection to practice scoring (with frequent comparisons to scoring by master coders).
- Ensure standardization with common training agendas and by training the trainers to use common slides and materials. Consider how a different training format than the one used in the first round of training may support consistency at scale (e.g., going from all in-person training to a hybrid of online and group work).
- Establish a process to collect feedback for improvement from observers and trainers on the quality of practice activities and tools. This should include identifying problematic videos to be removed from training (e.g., videos that prove to be distracting, represent difficult-to-judge borderline scores, or prompt debate).

CONTINUALLY IMPROVE

□ **Begin annual follow-up training to keep observers calibrated and to develop their skills to more sophisticated levels.**

Skills can rust, and observers need to know more than they can absorb in their initial training.

- Teach observers more efficient ways to record evidence (e.g., use codes for commonly observed behaviors, like "CFU" for "check for understanding").
- Focus follow-up training on less common situations that involve more nuanced distinctions among components and performance levels, such as when different aspects of the same criterion are observed at widely different performance levels (e.g., students showed a grasp of academic vocabulary, but the teacher did not).

ACTION STEPS TO BUILD OBSERVER TRAINING

MODELING FEEDBACK. *Illustrate how to give teachers productive feedback based on observations.*

LAY THE FOUNDATION

□ **Show observers how correct scoring and use of evidence supports effective feedback.**

Effective feedback won't happen if observers don't understand how to provide it or why it's essential.

- Clarify how good feedback and quality evaluation support fairness and trust (e.g., accuracy allows for a clear and common language, and supportive feedback makes it easier to accept critiques).
- Show videos of effective post-observation conferences and call out what makes them effective (e.g., the teacher does most of the talking, discussion is grounded in the rubric and objective evidence from the lesson, and the conference ends with agreement on one or two small changes the teacher will make before the next observation).
- Give observers a general agenda or protocol to follow in post-observation conferences.

BUILD STRUCTURES

□ **Require observers to practice giving feedback, and give them feedback on their feedback.**

Trust will be lacking if teachers across the system don't experience observations as positive and supportive.

- Have observers role-play mock post-observation conferences based on instruction seen in pre-scored videos.
- Have observers critique videos or role-play examples of effective and ineffective post-observation conferences.
- Begin an annual survey of teachers and ask to what extent post-observation conferences provided them with clear and actionable feedback.

CONTINUALLY IMPROVE

□ **Begin annual follow-up training on feedback that focuses on how to handle different situations.**

The most effective feedback is tailored to the learning style and needs of the one receiving it.

- Provide guidance on how to adjust feedback discussions depending on whether teachers are struggling, highly effective, defensive, or unreflective.
- Train observers on how to coach teachers through co-lesson planning, modeling, and role-playing.

Evaluate Whether Training Was Successful

Verification supports credibility. Making sure observers possess at least a minimum level of accuracy in scoring before they evaluate in the classroom builds confidence in the results—among teachers, among state and district leaders, and even among observers themselves. Moreover, observer assessment provides essential information with which to retrain individual observers and to plan overall improvements in a training program, which in turn drives improvements in accuracy. It's essential to know that observers can apply a rubric as intended.

Building an assessment process requires many of the same capacities needed to develop training. The same expert observers who pre-score videos for observer training need to pre-score videos that can be used to determine the extent to which observers who complete training are able to correctly score entire lessons.

It's essential to know that observers can apply a rubric as intended.

But assessment also presents a number of special challenges. One of the biggest is determining and communicating a standard for accuracy amid the recognition that everything about an observation system will improve over time. It's important to set a minimum threshold of accuracy for observers to demonstrate when stakes are attached to their evaluation of teaching. But observers will become more skilled as they get more experience with a rubric and as the training they receive is refined and enhanced.

Observer assessment is best planned for from early on in the implementation of an observation system. Assessing observers from the beginning signals a commitment to accuracy and sets the expectation among observers that they must demonstrate their proficiency at the end of their training. Although it represents a major adjustment for evaluators who have not previously been evaluated themselves this way, the expectation is more easily accepted if set from the start.

In reality, observer assessment may not yet have been developed in many places where the implementation of an observation system already is underway. In such contexts, a plan to evaluate observers against a set standard for

KEY TERMS

Observer assessment. The process of determining the extent to which observers are able to score correctly, typically by having them rate a set of videos that have been pre-scored by master coders.

Certification. The determination made through assessment at the end of initial training that an observer has at least a minimally sufficient level of accuracy.

Calibration. The periodic reassessment of observers to determine if they have maintained sufficient accuracy. Sometimes called *re-certification* or *norming.*

accuracy will need to be a top priority. If the result is that observers get more individualized data on their skills, they will be more likely to accept this new expectation. More importantly, they will be more likely to become better observers.

The sequence of action steps in the following pages lays the foundation for observer assessment by focusing first on using the results to build supports that help all observers meet a minimum standard of accuracy. Standards are then adjusted as more observer assessment data are available, as assessment tools and training improves, and as observers get used to the process. However, passing thresholds should be set not with the goal of certifying all observers, but with teachers and students in mind. Allowing observers to be far from accurate on many components of teaching opens the door to inconsistency, undermining the goals of fairness, instructional improvement, and trust.

ESSENTIAL ACTIVITIES FOR DEVELOPING OBSERVER ASSESSMENT

Action steps to address each activity are on the following pages.

Determining tasks. Create a scoring activity that mirrors what observers will do in the classroom.

Defining accuracy. Set a minimum standard for scoring proficiency.

Establishing consequences. Clarify what happens when observers fail to demonstrate sufficient accuracy.

An Example of DEFINING ACCURACY

Raters who participated in the MET project's study of classroom observation instruments had to demonstrate a minimum level of accuracy in scoring pre-scored lesson videos before they could rate lessons. The examples below apply the passing threshold the project used for the Framework for Teaching rubric to two hypothetical examples.

Passing Requirements: At least 50% exact match with correct score. No more than 25% discrepant scores (2 or more points from the correct score).

RUBRIC COMPONENT	CORRECT SCORE (1–4)	SCORE GIVEN OBSERVER 1	SCORE GIVEN OBSERVER 2
Creating an environment of respect and rapport	4	🟧 2	🟧 2
Establishing a culture of learning	3	🟩 3	2
Managing classroom procedures	3	🟩 3	🟩 3
Managing student behavior	4	3	🟧 2
Communicating with students	1	🟧 3	🟧 3
Using questioning and discussion techniques	2	🟩 2	🟩 2
Engaging students in learning	2	3	🟩 2
Using assessment in instruction	3	🟩 3	2
		Passed	Did Not Pass

🟩 Exact match
🟧 Discrepant score

Note: Assessment of observers in the project involved scoring multiple pre-scored videos. Examples shown are based on one lesson to clarify the general process.

ACTION STEPS TO DEVELOP OBSERVER ASSESSMENT

DETERMINING TASKS. *Create a scoring activity that mirrors what observers will do in the classroom.*

LAY THE FOUNDATION

□ At the end of an initial round of training, require observers to rate at least two pre-scored, observation-length videos from the range of subject areas and grade levels they will evaluate.

Confidence in observation comes from evidence that the observers are able to score correctly.

- Natural variation among lessons means one video is not enough; an observer might score one lesson correctly but another one incorrectly. (In such cases, a third video may be used to confirm the result.)
- Assessment videos should present clear-cut examples of different levels of practice, not borderline cases. They should capture typical practice and behaviors likely to happen in the vast majority of classrooms. Assessment tasks should also mirror the scoring process observers will use in the classroom (e.g., if they must provide scores for each component of teaching for each 30-minute lesson).

BUILD STRUCTURES

□ Replace assessment videos identified as problematic in the first round of observer assessment.

It may become clear only after initial use that some videos include borderline examples of scores and should be removed.

- No observer should score the same assessment video more than once; allowing for multiple scoring of the same video by the same observer compromises an assessment meant to determine accuracy in scoring lessons seen for the first time.
- Clarify to observers that they will be re-assessed and why. Explain how natural tendencies can affect accuracy over time.

CONTINUALLY IMPROVE

□ Begin re-assessment of observers (at least annually) with a plan in place to continually refresh the supply of videos available for assessment.

Follow-up training is not enough to ensure that observers have maintained the ability to score correctly.

- Re-assessment should meet the same minimum criteria for initial assessment: at least two videos, and mirror the process of scoring in the field.

DEFINING ACCURACY. *Set a minimum standard of scoring proficiency.*

LAY THE FOUNDATION

□ Set a minimum threshold for accuracy, primarily for the purpose of gathering data with which to improve supports for observer accuracy.

Assessing from early on clarifies to observers that there is a correct way to score and provides valuable information with which to ensure sufficient accuracy going forward.

- At first, all observers might not be expected to provide the exact correct score (determined by master coders) for most of a rubric's teaching components (e.g., for use of questioning or classroom management). But an early goal could be to at least get them to place the lesson as a whole in the correct category of performance. Component-level scores should still be collected to inform improvements in training and identify the most accurate observers for future master coding. While such a minimum standard may be an improvement over past observation practice in a system, it is not sufficient for ongoing implementation.

BUILD STRUCTURES

□ Consider results from the first round of observer assessment to set a passing threshold that specifies the extent to which observers' component-level scores must be correct.

Observers' scores for each observation should be accurate for at least the majority of teaching components.

- What's needed is a generally accepted minimum level of accuracy for credible evaluation and feedback. The amount of correct component-level scores that may be expected depends on the range of possible scores (e.g., fewer exact matches on a 7-point rubric than a 4-point one).
- For a 4-point rubric, observers' component-level scores should exactly match the correct scores determined by master coders at least 50 percent of the time to be considered proficient. If the required exact-match rate is kept to just 50 percent, then establish the extent to which observers' scores may be more than one point away from the correct score.

CONTINUALLY IMPROVE

□ Establish an ongoing process to consider adjustments to how passing thresholds are defined based on passing rates and feedback from observers and teachers.

A more refined understanding of what's minimally acceptable for accuracy will emerge over time.

- Passing thresholds should not be lowered simply to increase the number of observers who pass. Doing so compromises the quality of evaluation. If not enough observers are able to demonstrate a minimally acceptable level of proficiency, then the right response is to improve training.
- A higher threshold of accuracy may be set to identify the most accurate observers as master coder candidates.

ACTION STEPS TO DEVELOP OBSERVER ASSESSMENT

ESTABLISHING CONSEQUENCES. *Clarify what happens when observers fail to demonstrate a minimum level of accuracy.*

LAY THE FOUNDATION

□ **Prioritize observers who do not initially meet the minimum threshold for accuracy for additional training and support.**

Retraining increases the quality of results and further gives observers the chance to develop what may be a new skill for them before more consequences are introduced.

- Ensure that observers whose scores are the furthest off from the correct scores are prioritized for the most intensive support. Consider having struggling observers observe alongside observers who have demonstrated a high level of accuracy.
- Communicate that a passing threshold will be defined after this initial round and may change over time as training evolves and results are analyzed. Use assessment result trends to identify components of the rubric for which training may need to be strengthened.

BUILD STRUCTURES

□ **Communicate to observers that they cannot evaluate teachers for stakes until they demonstrate sufficient accuracy, and continue to retrain and support those who don't meet the minimum threshold.**

Trustworthy observations depend on accurate observers.

- Keep in mind that some individuals may not be able to evaluate teachers because they cannot demonstrate proficiency even after retraining. Observers who fail certification may observe in the classroom alongside accurate observers who provide the official scores.
- Keep in mind that even if an individual's assessment results are not made public, it may become known that an administrator is not permitted to carry out official observations in the classroom. Districts may need to go beyond the traditional pool of observers to ensure that all teachers receive multiple observations that can be averaged to produce reliable summative evaluations.

CONTINUALLY IMPROVE

□ **Continue a policy of certifying observer accuracy, using assessment results and input from trainers to further target retraining to individuals' specific needs.**

Differentiated support is more likely to get observers over the bar.

- Consider different categories of certification (e.g., for evidence collection, alignment of evidence to rubric component, and accurate scoring) to direct retraining to more specific skill sets.
- Consider using certification assessment results to identify high scorers to assist in coaching observers who initially do not pass and to play other leadership roles.
- As with all assessments, results should be analyzed to make sure observer assessments are not unfairly biased against individuals with certain backgrounds.

Check that the System Is Working As Intended

The only way to know that an observation system is functioning as intended is to check. Without ongoing monitoring, natural tendencies to deviate from expectations will go unnoticed and unaddressed. Moreover, it is through data gathered at the system level that a state or district is able to evaluate the return on investments in improved instruction, allowing for better decisions about professional development and policy. Accurate observations provide critical insights into the state of teaching within a state or district, without which significant improvement is unlikely.

Ongoing monitoring also supports fairness. If some observers don't follow procedures, then the teachers whose lessons they score are less likely to benefit from accurate feedback and evaluation. Even observers who follow procedures may be unaware that their scoring has drifted since the time of their initial training, resulting in scores for teachers that are unacceptably different than what master coders would give. Some observers able to score accurately in a video-based assessment may nonetheless struggle to set aside prior knowledge of teachers in their schools and focus solely on the lesson they see.

Without ongoing monitoring, natural tendencies to deviate from expectations will go unnoticed and unaddressed.

The sequence of action steps in the following pages builds from a set of procedures to set expectations and audit accuracy to the use of information to better target supports for teachers and observers. The foundational step is to make sure that observations are happening—a seemingly obvious expectation but one easily lost amid competing demands on people's time. (Indeed, making time for evaluators to observe is essential to trustworthy observations.)

Another important aspect of monitoring is to look for patterns that may indicate the need for additional training of observers. An observer might need additional training if he or she submits scores that are much higher or lower than those given by other evaluators, or if the observer's scores bear no relationship to other measures of effectiveness for the same teachers. Such observers may benefit from co-scoring with another observer

KEY TERMS

Drift. A tendency to inflate or deflate scores over time. A single observer may drift, or a group of observers may drift in the same direction.

Double scoring. The process of assessing the overall reliability of an observation system by having some teachers observed by more than one observer.

Aberrant scores. Scores that show significantly different patterns than the norm and that may be a sign of inaccuracy.

to re-norm their application of the rubric. Along with observation scores and reports, teachers and observers are important sources of information about how well an observation system is working.

The ultimate measure of the success of an observation system is the extent to which teaching becomes more effective. It's important to keep in mind that this takes time—for observers to practice, for teachers to make small adjustments in their instruction, and for the effects to show up in better student outcomes. States and districts that have stuck with the process the longest are starting to see the benefits.

ESSENTIAL ACTIVITIES FOR CREATING OBSERVATION MONITORING

Action steps to address each activity are on the following pages.

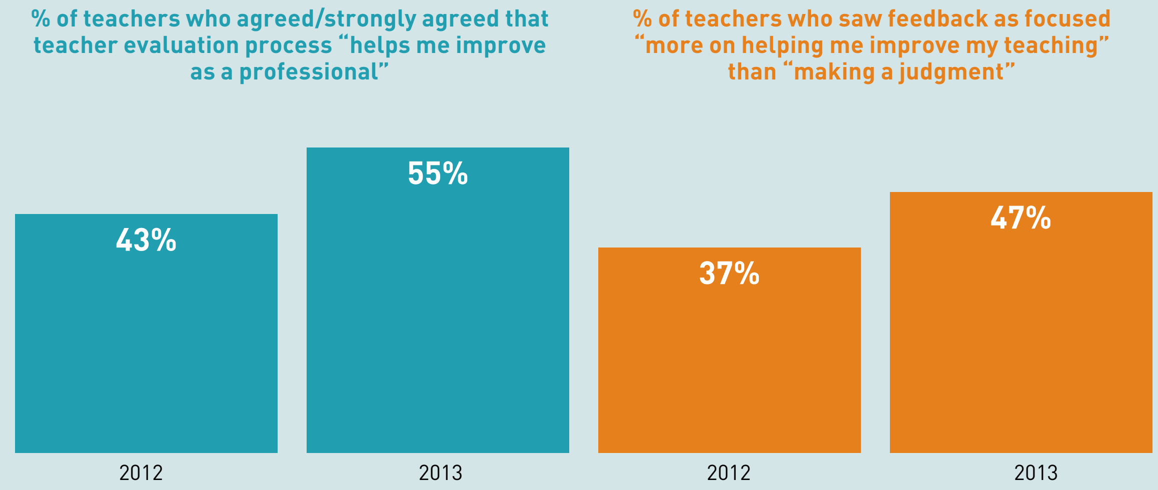
Verifying process. Inspect to see if observation procedures are followed.

Checking agreement. Make sure observers maintain their accuracy.

Evaluating support. Assess efforts to improve instruction.

An Example of **EVALUATING SUPPORT**

Annual surveys are used to gauge changes in how Tennessee teachers see their state's evaluation system, in which observations play a major role.



Source: Tennessee Consortium for Research, Evaluation, and Development.

ACTION STEPS TO CREATE OBSERVATION MONITORING

VERIFYING PROCESS. *Inspect to see whether observation procedures are followed.*

LAY THE FOUNDATION

□ **Clarify observation procedures to evaluators and establish a system for submitting information from observations.**

Accurate and fair observations depend on consistently followed procedures.

- Make sure observers know how to complete observation forms (what parts are required and if any are optional), how often they must observe each teacher, for how long, and by what dates.
- Set up an online system for observers to submit observer scores, teacher names, dates, and other information. Ensure that observers know how to use this system and when they must submit what information.

BUILD STRUCTURES

□ **Assign ongoing responsibility for determining whether observations are being carried out according to procedures.**

Observations won't take place according to procedures if someone isn't checking to make sure.

- Institute training and guidelines to address specific compliance issues identified during initial implementation.
- Review submission of observation information regularly and follow up with any observers for whom items are missing.

CONTINUALLY IMPROVE

□ **Continue monitoring adherence to observation procedures, including use of teacher surveys.**

Additional verification heightens accountability and better identifies problems.

- Ask teachers if observations are taking place according to schedule, and if observers are following specified procedures for any pre- and post-conferences. When observers who fall behind schedule must catch up at the end of the year, the observation process can become one of compliance rather than professional development.

CHECKING AGREEMENT. *Make sure observers maintain their accuracy.*

LAY THE FOUNDATION

□ **Begin using observers' attempts to score master-coded video to target retraining and to address areas of the rubric where many observers struggle.**

When observers are new to scoring with a rubric, the best way to monitor their scoring ability is with examples for which the correct scores have been carefully determined.

- Data on observer accuracy may come from assessment at the end of training but also from short, low- or no-stakes calibration scoring activities assigned throughout the school year. (For more on master coding, see action steps for pre-scoring video, page 16.)

BUILD STRUCTURES

□ **Start identifying places where teachers' observation scores and student learning measures show widely different patterns as possible instances of aberrant scoring.**

Patterns may suggest grade inflation, tendencies to score too low, or a misunderstanding of the rubric.

- Provide reports to school and district leaders that compare their score distributions to those of the larger system.
- Create a plan to respond to aberrant scoring with additional monitoring and support. This might involve sending in expert observers to a school or district to observe alongside local evaluators and to provide coaching to norm their application of the rubric.

CONTINUALLY IMPROVE

□ **Establish a process in which some of the same lessons may be scored by more than one observer to monitor observer agreement.**

If different observers score the same lesson correctly they should produce the same scores.

- Double scoring is most productive when observers have developed a foundation of scoring proficiency. If one of the observers is known to be accurate, then the other can benefit from the comparison.
- If it is logistically difficult to have more than one observer in the classroom at the same time, then video may be used to allow for independent scoring.

ACTION STEPS TO CREATE OBSERVATION MONITORING

EVALUATING SUPPORT. *Assess efforts to improve instruction.*

LAY THE FOUNDATION

□ **Communicate to all stakeholders the intent to use observation data to strengthen supports for improved instruction.**

Teacher buy-in depends on the understanding that teachers will be better served as a result of an observation system.

- Make sure all leaders across the system convey a consistent message in all communication about the observation system.
- Point out ways in which observation data will be used to improve supports (e.g., determining and evaluating professional development or evaluating and training school leaders).

BUILD STRUCTURES

□ **Use observation data to inform systemwide decisions about investments in professional development.**

Teachers should get what they need, not what they don't.

- Shift teacher training resources to areas of greatest need and with the greatest potential to improve student outcomes (e.g., if most teachers are good at classroom management, then invest less in classroom management training and more in professional learning opportunities that can help teachers move more students to higher levels of academic mastery).
- Establish a process to use observation data in evaluating which supports and interventions work best for specific populations of teachers (e.g., new teachers versus teachers rated on the cusp of being highly effective).

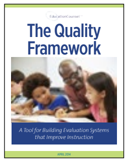
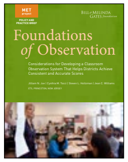
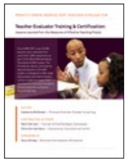
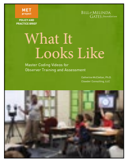
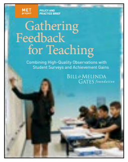
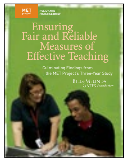
CONTINUALLY IMPROVE

□ **Use the extent to which teachers move to higher levels of performance to evaluate professional development, school and system leadership, and policy decisions.**

The potential for improvement increases dramatically if all parts of a school system are driven by data on the quality of instruction.

- Stop investing in professional development that doesn't result in improved practice.
- Identify, celebrate, and learn from places in the school system where teachers are consistently elevating their levels of practice.

Additional Resources

	Title	Source	Content
	**The Quality Framework: A Tool for Building Evaluation Systems that Improve Instruction (2014)**	Education Counsel	A resource to help state education leaders plan and improve an evaluation system based on multiple measures. Provides implementation criteria, a self-assessment, and suggestions on where to find additional guidance on specific implementation issues regarding observations, as well as other measures.
	**Foundations of Observation (2013)**	The MET project	A white paper by experts at ETS on the elements of observer training and assessment that can produce accurate and reliable results for teachers.
	**Teacher Evaluator Training and Certification (2012)**	Teachscape	A paper from a lead partner on the MET project on the use of video to train observers and certify their accuracy.
	**What It Looks Like: Master Coding Videos for Observer Training and Assessment (2013)**	The MET project	A paper by a lead architect of the MET project's observer training and assessment system that details a model for pre-scoring, or master coding, video to build and maintain observer accuracy.
	**Gathering Feedback for Teaching (2012)**	The MET project	A policy and practice brief that explains research findings on the reliability and validity of classroom observations. A longer research paper of the same name includes details on the study's training, assessment, and monitoring of observers.
	**Ensuring Fair and Reliable Measures of Effective Teaching (2013)**	The MET project	A policy and practice brief on three major MET project studies that includes discussion of how multiple observations can ensure reliable results.
	SOE Teaching and Learning Exploratory (TLE)	University of Michigan School of Education	A website of authentic teaching videos, lesson materials, and interactive tools available to individual and group subscribers. Includes multiple collections of teaching videos, including more than 1,500 recorded as a part of the Measures of Effective Teaching Extension project. By special arrangement, individuals and institutions can upload their own videos into TLE private channels.