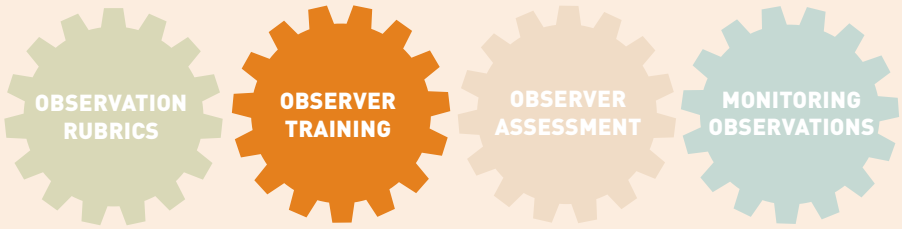
MAKING IT REAL

Pre-Scoring Video To Clarify Expectations For Effective Teaching



CONTENTS

November 2014

BILL & MELINDA GATES foundation

Picture this scene: As part of their training, a group of evaluators are asked to rate a video of instruction using their district's observation rubric. They each review the clip, take detailed notes, and seek to align what they saw and heard with the right level of performance. But the discussion quickly turns into debate over how to rate the teaching on a key component in the rubric.

One camp says the video showed *proficient* performance for "use of questioning." Another says *needs improvement.* Drawing on years of experience that have shaped their visions of effective instruction, both camps cite evidence they feel is most relevant to their interpretation of the teacher's practice. Despite extended conversation, no clear consensus is reached when the session ends.

Learning to assign the right score for the right reasons isn't just about accuracy. It allows evaluators to clearly explain *this is your score, this is why, and this is how to improve it.*

This scenario, which likely echoes similar ones in many school systems, poses some major problems. The evaluators leave with significantly different interpretations of the same practice, undermining the fairness of observations and the quality of the data they produce. The resulting inconsistencies will soon be apparent to teachers. The evaluators themselves may be frustrated by not having had the opportunity to learn how to accurately rate the video. They haven't been set up to provide consistent and specific feedback.

Fortunately, there's a way to avoid these outcomes. To build a common understanding of what it looks like when teaching is demonstrated at particular levels takes practice with examples of teaching for which accurate ratings have been determined. This is why pre-scored video is foundational to observer training. Pre-scored video is used to clarify parts of a rubric, to practice evidence collection and scoring, and to assess observer proficiency. Learning to assign the right score for the right reasons isn't just about accuracy. It allows evaluators to clearly explain *this is your score, this is why, and this is how to improve it*. Giving teachers access to pre-scored video also helps them better understand a school system's expectations and demystifies the observation process.

A FOUNDATIONAL PROCESS

Pre-scoring video serves multiple purposes in an observation system:

- To inform development of a new rubric, or revisions to an existing one
- To identify challenges to target in observer training
- To give concrete examples of a rubric's components, performance levels, and terms
- To let observers compare their own attempts to rate with accurate ratings
- To assess observer accuracy at the end of training and on a periodic basis
- To build a cadre of experts in a rubric who can play leadership roles in a system

States or districts that adopt an existing rubric may use training that includes already pre-scored video, but an understanding of pre-scoring is important in any trustworthy observation system. Where a rubric is new or modified (e.g., to align with new college- and career-ready standards), pre-scoring should be part of the instrument's development; attempting to score actual lessons consistently is an important test of applicability. Commercially available observer training also may need augmenting with examples from the local context. Moreover, the pre-scoring process—in which individuals reach agreement on how to accurately rate videos—can be a powerful professional learning opportunity for teachers and instructional leaders, sharpening a shared vision of effective teaching.

Using this guide: This guide can help a school system build and strengthen a pre-scoring process in which expert observers ("master coders") create examples of accurate and well-supported ratings of teaching. Along with tools, tips, and snapshots of practice, each section includes techniques and idea-generating prompts for starting a process (under "To Lay the Foundation") and for improving one (under "To Build and Improve"). Use the worksheet of essential questions at the end to summarize plans.

This is one in a series of MET project practice guides for states, districts, and technical assistance providers on how to build and improve a trustworthy observation system. To assess an observation system's status and plan for continual improvement of all its components, see *Building Trust in Observations: A Blueprint for Improving Systems to Support Great Teaching.* All MET project resources are at: www.metproject.org.

What Is "Pre-Scoring Video"?

Pre-scoring video is the anchoring of an observation rubric to actual examples of teaching. This is much the same as expert reviewers scoring and annotating examples of students' written responses to anchor a common understanding of how to evaluate answers to open-ended questions on a standardized test. In pre-scoring video, segments of recorded lessons are anchored to the right performance rating for each teaching component the segment illustrates. (Other terms for pre-scoring include "anchor rating," "master coding," and "master scoring.") Pre-scored video makes concrete the written definitions in a rubric.

But more than just the right rating is needed for pre-scored video to serve its purpose. The correct justification for that rating is equally important. When evaluators understand what behaviors observed in recorded lessons indicate a particular rating, they're better able to rate other lessons accurately and in a variety of contexts. In addition, when they understand why a different rating would not be correct, they can more clearly communicate to teachers what would result in a higher rating. This supports both accuracy and actionable feedback. For this reason, a quality pre-scoring process produces clear justifications not just for why a rating is correct, but also for why other ratings would not be.

The example in **Figure 1** below shows the key elements of well-supported ratings for a pre-scored video. The rationale justifies a rating of *effective* for "use of questioning" in a four-minute segment, or excerpt, from a lesson video. The explanation of the correct rating, and why other ratings would be incorrect, is grounded in the language of the rubric. Key indicators are the extent to which questions invite student thinking and ask students to explain their reasoning. Exact quotes of the teachers' questions, with timestamps for when in the video they were asked, show that although the questions invited student thinking and discussion, none of them encouraged students to demonstrate their reasoning.

FIGURE 1. **WELL-SUPPORTED RATINGS FOR A PRE-SCORED VIDEO SEGMENT**

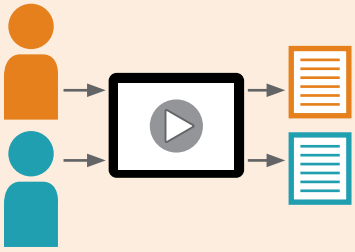
USE OF QUESTIONING: Effective

Evidence	Score Rationale
Teacher questions: ■ 14:02 "What tools would a scientist use?" ■ 16:58 "What would a butterfly do?" ■ 17:59 "How is the pollen going to come off the flower and go to another?"	■ **Why the rating is *effective.*** Most of the questions the teacher asks are open in nature and engage students in deeper thinking and further discussion. ■ **Why a lower score is wrong.** The teacher does not use a combination of open and closed questions, with only some questions inviting thoughtful response. ■ **Why a higher score is wrong.** The teacher's questions do not provide students an opportunity to demonstrate reasoning for formulating their own questions.

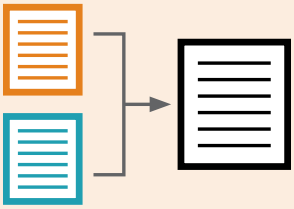
Pre-scored video sets the gold standard for accurate rating with a rubric. Given that, it's important to make every effort to ensure that the ratings and justifications produced are both clear and correct, based on the rubric. If not, their use in training will not support observers in rating both consistently and accurately. For this reason, pre-scoring involves multiple expert observers, or coders, who score the same segments individually, after which they compare notes to reach agreement—a process called reconciliation. (See **Figure 2** on the next page.) Without this process, a single master coder may overlook relevant evidence or fail to notice a key difference in how a rubric distinguishes between two performance levels.

FIGURE 2. STEPS IN PRE-SCORING VIDEO

1. Expert observers independently review video segments and submit score rationales based on the rubric.



2. Submissions are compared and differences reconciled to produce a single set of scores and rationales.



3. Video segments are used in training with reconciled scores and rationales to align trainees' understanding of the rubric.

But success in pre-scoring video also depends on many other factors, such as the expertise of the master coders and their understanding of the pre-scoring process, the quality of the video to be scored, the specificity and language of written score justifications, and the use of data for continual improvement of the process. Strategic planning also is needed to build support for the process, to identify the required resources, and to build the capacities in a way that best meets a school system's ongoing needs. Each of these issues is addressed in the sections that follow, with advice on how to start a pre-scoring process and how to strengthen it over time.

SNAPSHOT
PRE-SCORING VIDEO AND THE AMERICAN FEDERATION OF TEACHERS

The American Federation of Teachers includes pre-scoring video as part of teacher evaluation work funded by a U.S. Department of Education Investing in Innovation (i3) grant. State affiliates in New York and Rhode Island have worked with a dozen districts to pre-score video to support teacher and evaluator training aligned to rubrics shared across the participating school systems within each state. Taking part in the pre-scoring process are teacher leaders, peer evaluators, principals, and central office administrators.

"It makes you think about the rubric so much more deeply—which makes you think about practice so much more deeply. It allows you to verbalize expectations and make it real for folks." *—Katrina Pillay, a participant from Cranston Public Schools in a pre-scoring process organized by the Rhode Island Federation of Teachers and Health Professionals*

"It improved our rubric. There were pieces where we felt the language needed to be fine-tuned. When you actually start using the rubric is when you say, 'I really don't know how to find this.' " *—Robin Hecht, a participant from the Marlboro Central School District in a pre-scoring process organized by the New York State United Teachers*

PUTTING IT INTO PRACTICE

TO LAY THE FOUNDATION:

Which parts of the pre-scoring process do you have the most questions about?

TO BUILD AND IMPROVE:

In which parts of your pre-scoring process do you see the greatest need for improvement or refinement?

See page 33 for guidance on using data to answer this question.

*Note: Saving notes in the fields above requires a recent version of **Adobe Reader**.*

How Will You Create Support For Pre-Scoring Video?

The investment required to pre-score video is not insignificant. It takes time to prepare video for scoring; to train master coders; and to review segments repeatedly to record evidence, align it to the correct parts of a rubric, and come to agreement on ratings. It also requires an ongoing time commitment. Typically, master coders are teachers, principals, and other instructional leaders who may devote several hours a month to pre-scoring throughout the year. Moreover, it may take up to two years before the process runs smoothly and consistently yields high-quality examples of well-justified ratings. Such an effort will succeed only if stakeholders appreciate its value.

In many places, classroom observations count for half or more of a teacher's overall evaluation. Trust in the whole evaluation system suffers when different observers apply the same rubric differently; teachers are left knowing that their ratings depend more on who does the observing than on how their teaching compares with a set of common expectations. But trust also falters if observations fail to produce actionable feedback. Absent clear, evidence-based explanations, even accurate performance ratings are little help to teachers who want to improve their practice. For evaluation to improve teaching and learning, observers must be skilled at providing well-supported performance ratings.

To develop this skill, video is almost essential. The ability to provide ratings that are both accurate and well supported comes from understanding why a particular instance of teaching merits a particular rating and from comparing one's own attempts to rate with accurate examples of rating the same instruction. This is very hard to accomplish with live observations. In a live observation with one or more observers, a single trainer must determine the correct rating on the fly. Classroom space also severely limits the number of evaluators who can observe the same live lesson. In contrast, video examples allow for:

- pre-scoring by multiple expert observers to confirm the correct rating and rationale;
- unpacking all the relevant evidence in a few minutes of teaching;
- rewinding and reviewing an instance of teaching multiple times; and
- exposing all observers to the same examples, regardless of where and when training takes place.

Trust in the whole evaluation system suffers when different observers apply the same rubric differently. And absent clear, evidence-based explanations, even accurate performance ratings are little help to teachers who want to improve their practice.

How support is built for pre-scoring video will evolve over time. When starting out, a state or district has few results to point to in making the case for this investment. To lay a strong foundation of support requires educating stakeholders about the importance of consistency and actionable feedback and about how both are advanced by examples of well-supported performance ratings. Later, a state or district can point to benefits experienced to build on that foundation. Participants in the pre-scoring process can be especially effective in communicating its value. Creating support means building trust among stakeholders in the pre-scoring process, in the master coders, and in what they produce.

SNAPSHOT
MAKING THE CASE FOR PRE-SCORING VIDEO IN COLORADO

The Colorado Department of Education created a set of summary documents to explain how and why it manages a master-scoring process. A short **fact sheet** explains who serves as "master scorers," why consistency in evaluation is important, and how the state is making pre-scored video available to teachers and school leaders through an online platform, Elevate Colorado. The two-page document states: "This process builds clarity in understanding and identifying high-quality teaching and thus consistency in evaluation."

For more on Elevate Colorado, see **www.cde.state.co.us/educatoreffectiveness**.

COSTS OF PRE-SCORING IN COLORADO

No one formula can determine the cost of pre-scoring video in every context. The biggest factors are the time for people to manage the process and the time of master coders, who are typically paid a stipend. **Figure 3** below, from the Colorado Department of Education, suggests the cost in dollars and time for each of the state's master scorers. In its first year of pre-scoring, Colorado trained and worked with six master scorers. Now with 13 master scorers, the state reviews and pre-scores about 12 videos a year. The time required for review and reconciliation may vary significantly among states and districts depending on the complexity and specificity of the rubrics.

Additional expenses include the time of department personnel who recruit and train master scorers, plus travel expenses related to training and work sessions. The cost of getting video of teaching to pre-score depends on the source. Colorado has both purchased access to lesson videos from commercial vendors and contracts with a videographer to record new lessons.

FIGURE 3. COSTS PER MASTER SCORER IN COLORADO

Stipends	Time Commitments	
$2,750	Spring and summer training to learn process and begin pre-scoring ■ 3 days ■ Plus pre-work (includes reviewing video)	Eight work sessions during year to compare and reconcile ratings ■ Up to 3 hours for each session ■ Plus pre-work (1.5–3 hours to review video)

TECHNIQUES
BUILDING SUPPORT FOR PRE-SCORING VIDEO

To Lay the Foundation	To Build and Improve
■ Craft a concise message for funding proposals, information sheets, and recruitment efforts about how pre-scoring video supports consistent evaluation and actionable feedback. ■ Build credibility for pre-scoring video among teacher and administration groups by explaining how multiple instructional experts work together in a structured process to pre-score each video. Share how important this work is to building a consistent vision for effective instruction across the system. ■ Ask groups of educators not trained with pre-scored video to score part of a video and compare the ratings. Use examples of inconsistency to make the case to the group for pre-scoring. ■ Cite research, including studies by MET project partners, showing that observers trained and assessed with pre-scored video can produce consistent performance ratings. (See ***Gathering Feedback for Teaching*** research report.)	■ Share testimonials from teachers and administrators about how working as master coders has improved their own teaching and instructional leadership. Have them talk about the rigor of the pre-scoring process. ■ When using pre-scored video in training with evaluators new to the process, briefly explain how the correct ratings were determined by multiple instructional experts. ■ Communicate the professional backgrounds of master coders and how representative they are (from different grade levels, regions of the state or district, etc.). ■ Communicate examples of how the pre-scoring of video has improved parts of the observation system (e.g., identifying rubric components to clarify). ■ Share observer assessment data from your own state or district on the amount of consistency among observers trained with pre-scored video.

PUTTING IT INTO PRACTICE

TO LAY THE FOUNDATION:

What case for pre-scoring would be most compelling to multiple perspectives (teachers, school leaders, administrators) in your school system?

TO BUILD AND IMPROVE:

How can you more effectively communicate the benefits of pre-scoring video to educators across your system?

Note: Saving notes in the fields above requires a recent version of ***Adobe Reader****.*

What Is Your Goal for Pre-Scoring Video This Year?

Pre-scoring video is not a one-time activity. It's an ongoing process in a state or district's efforts to foster a shared understanding of effective teaching. It's also a capacity that can be developed only through practice. With each round of pre-scoring, participants learn better how to produce examples of well-supported performance ratings that can help evaluators do the same. Understanding this evolution helps with near-term planning. Goals for pre-scoring in any given year should be determined based on how the process can best add value to current efforts. That changes as the capacity to pre-score video increases and as an observation system continually improves.

With each round of pre-scoring, participants learn better how to produce examples of well-supported performance ratings that can help evaluators do the same.

Ideally, the first time videos are scored with a rubric is during rubric development. Doing so "pressure tests" the way the rubric distinguishes among different teaching components and performance levels. Indeed, trying to get a rubric exactly right before attempting to score lessons with it can be a waste of time. Only when evaluators try to apply a draft instrument to actual teaching will it become apparent where wordsmithing—or more substantive changes—may be needed. Using video makes it easier for developers to understand why different observers might struggle to apply the instrument. If pre-scoring video was not part of rubric development, then test the rubric with some video scoring before proceeding.

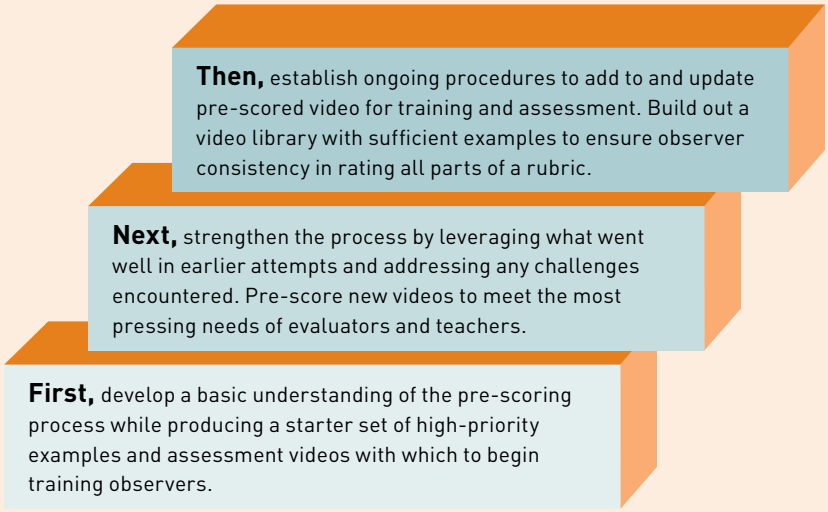
If there's evidence that evaluators can apply a rubric, a group of master coders can be trained to produce pre-scored videos for observer training. The training of master coders is discussed later in this guide, but a key feature is the use of existing pre-scored video to build an understanding of the process. When there are no trained master coders yet to produce these videos, then the task must fall to those who have the best understanding of the rubric. This also pressure tests a pre-scoring process before teaching it to others. (At least one of the rubric experts who pre-scores video to train master coders should also lead the master coder training.)

It's unrealistic to expect to pre-score quality examples for most of a rubric in a matter of months. It's also not advisable. A school system's second attempt at pre-scoring is more likely than its first to produce accurate and well-supported ratings—and its third more likely than its second. Moreover, if a rubric is likely to change (e.g., to better align with the Common Core State Standards), then different examples will be needed. Caution is advised in revising rubrics, as even small changes in wording can alter interpretation. But if a rubric changes, then previously pre-scored video may need re-scoring. Better to learn the process while pre-scoring a limited number of videos than to do extensive rebuilding later.

When a solid understanding of the process is in place, the focus can shift to ensuring continual improvement while building out a video library. (See **Figure 4** on the next page.) At this point, ongoing procedures are needed to replace outdated or problematic videos, to refresh the supply of assessment videos, and to prioritize additional parts of a rubric to illustrate with new examples. The pre-scoring process itself should continue to improve as more becomes known about the quality of examples produced and the effectiveness of their use in observer training. Pre-scoring video never ends in a trustworthy system of classroom observations. Along with the need for additional and updated examples for use in training, the need for a community of educators to better understand effective teaching is perennial.

FIGURE 4. BUILDING CAPACITY TO PRE-SCORE VIDEO

Building capacity to pre-score video is a matter of learning by doing. With each attempt, a school system learns better how to produce accurate examples of well-supported performance ratings. Starting small allows for developing a basic understanding of the process before establishing procedures to build out, update, and continually improve a more extensive video library. For guidance on building capacity in all parts of an observation system, see ***Building Trust in Observations: A Blueprint for Improving Systems to Support Great Teaching***.



TECHNIQUES
SETTING NEAR-TERM GOALS FOR PRE-SCORING VIDEO

To Lay the Foundation	To Build and Improve
■ To support a rubric's initial development, ask evaluators to try using the instrument to rate lesson videos. This may suggest where the rubric needs clarifying (e.g., if the rubric doesn't sufficiently define "students are generally engaged").	■ Establish an ongoing process for pre-scoring new video. Additional examples are always needed to build out and update a video library that supports observer training.
■ After a rubric is developed, have a small group of experts who know the instrument best pre-score three to four lesson videos to use in training a first cohort of master coders.	■ Replace pre-scored videos that proved to be problematic in training (e.g., if something in them distracted evaluators or they could not see or hear sufficient evidence to rate).
■ Train a first cohort of master coders to pre-score a small number of videos that can be used in the first iteration of observer training.	■ Replace videos used for assessment so that evaluators do not see the same one multiple times.
	■ When parts of a rubric change, ask evaluators to score lesson videos to make sure the new language is interpreted as intended. Rubric revisions may require new examples of well-supported ratings.

PUTTING IT INTO PRACTICE

TO LAY THE FOUNDATION:

What do you think are realistic goals for your first year of pre-scoring video?

TO BUILD AND IMPROVE:

Which of the techniques for building and improving on the previous page most resonate as goals for your system?

*Note: Saving notes in the fields above requires a recent version of **Adobe Reader**.*

What Pre-Scored Video Do You Need?

Different parts of a job call for different tools. A video segment for illustrating part of a rubric requires different content than one for practicing to rate entire lessons. Understanding which pre-scored videos can best add value to an observation system at a particular stage of implementation increases the chance of producing what's needed. Granted, pre-scoring video is a lot like mining: You can't know what you'll find before you start digging. You can't tell if a video shows what you need until you review it. But knowing what you need comes first. How to pre-screen video and identify segments for different uses is discussed later in this guide. Here, the guidance is on setting goals for what to produce.

At any stage of implementation, two categories of pre-scored video are needed: short training videos that show what parts of a rubric look like in practice and longer assessment videos that allow observers to score a lesson on multiple rubric components. This supports observer training that builds from the scoring of parts of a lesson on individual teaching components to the scoring of entire lessons, as observers do in the classroom. After training, at least two assessment videos are needed to meaningfully measure an observer's accuracy (additional videos may be required when the results from rating two videos appear to conflict.)

The previous section made the case for starting small. When pre-scoring for the first time, focus on building a basic understanding of the process while producing a limited starter set of pre-scored videos with which to begin training observers. The most useful training videos at this point will be short segments that each feature a clear-cut example—called a benchmark—of one teaching component demonstrated at one performance level. A benchmark might be a four-minute segment that shows "highly effective use of questioning" or a seven-minute clip showing "basic management of classroom procedures"—each annotated with a rationale calling out clear evidence to support the rating. A first set of assessment videos also should be produced in time to gauge observers' accuracy when they complete the new training; this provides valuable data with which to improve the training's next iteration.

RULES OF THUMB FOR PRIORITIZING

Only so many videos can be pre-scored in any period of time. To prioritize goals for the near term, ask: "Where is the greatest need for consistency in ratings?" and "Where is the greatest opportunity to improve teaching and learning across the school system?"

Consider how most teachers currently perform and where observers need the most help to better rate performance. For example, most teachers typically perform in the middle performance categories, and yet those are often hardest to distinguish with consistency. If this is the case in your system, then helping observers to better discern among 2s and 3s on a four-point rubric may result in more accurate and specific feedback for the most teachers.

Use data to identify where greater consistency or clarity on practice will be most beneficial. Before a pre-scoring process begins, ask a group of evaluators to rate a set of unscored lesson videos to see which parts of a rubric produce the most inconsistency. After a pre-scoring process is established, look at observer assessment results and teacher observation scores to determine new priorities.

Over time, a video library should be built out with enough examples to allow for sufficient observer agreement on all parts of a rubric.

After an initial set of videos has been used in training—and a basic understanding of pre-scoring is established—a state or district can turn its focus to building out its set of examples, to producing more videos to use in practice rating, and to replenishing its supply of assessment videos. Most of the training videos produced will likely continue to be benchmark examples. These may be of additional parts of a rubric, multiple examples of the same parts of a rubric, or replacements of previously pre-scored videos that proved to be problematic when used in training. Variations can address more specific needs, such as clarifying the meaning of a single term in a rubric (e.g., what is meant by "scaffolding"). Borderline examples—called rangefinders instead of benchmarks—can help clarify when teaching is at the low or high end of a rating (e.g., when the evidence of "student engagement" minimally qualifies for an effective rating and not the next lowest rating).

From the outset, it's important to keep in mind that teaching looks different in different places. Evaluators who are going to observe high school teachers should be trained and assessed with videos from high school classrooms. If observers will be rating math instruction, their preparation should include video of math lessons. Moreover, classrooms from the same grade and subject look different in different parts of a school system. The backgrounds of students and teachers differ, facilities differ, and teachers' personalities differ. Any set of videos used for observer training should reflect these differences. Evaluators must be able to recognize a rubric's indicators in any context they find themselves.

TECHNIQUES
PRIORITIZING PRE-SCORED VIDEO TO PRODUCE

	To Lay the Foundation	To Build and Improve
Training Videos. Short segments that illustrate one part of a rubric	■ Pre-score a small number of clear-cut examples (benchmarks) that each illustrate one teaching component at one performance level (e.g., a level 3 for "maximizing use of instructional time"). Often 2–12 minutes long. ■ Consider prioritizing what examples to look for first based on how most teachers in the school system perform and what parts of the rubric are hardest for evaluators to rate. To determine the latter, ask evaluators to rate unscored videos and look for parts of the rubric that produced the greatest inconsistency. (Keep in mind that locating examples of teaching components that show up less often will require reviewing more lessons to find them. See *How Much Raw Video Is Needed?* on page 14.) ■ From the beginning, make sure any set of videos used for observer training covers a range of grades, subjects, classroom compositions, and teacher backgrounds.	■ Pre-score additional benchmark videos for missing parts of the rubric, and replace existing examples with better ones. Consider where multiple examples of the same part of a rubric are needed to clarify how it might look in different classrooms, grades, subjects, etc. ■ Consider where rangefinder examples are needed to clarify boundaries between two ratings (e.g., the difference between a low 3 and a high 2). ■ Pre-score mini-segments—as short as one minute—that each illustrate key terms in the rubric that may be new to observers or that may be best defined through examples of practice. ■ Look to teacher observation results and observer assessment data to determine what additional examples to prioritize for pre-scoring. Ask what examples will most likely support better feedback to improve practice for the most teachers.
Assessment Videos. Longer videos used to determine the extent to which observers can score correctly	■ Pre-score segments that show clear examples of multiple, or all, teaching components in a rubric and that are similar in length to the observations to be carried out in the classroom—often approximately 20 minutes. ■ Pre-score at least two such assessment videos for each grade band in which observers will be evaluating. So if observers will be rating performance in two grade bands (e.g., K–8 and 9–12), then at least four assessment videos are needed.	■ Establish an ongoing process to pre-score additional assessment videos to replace problematic ones and to replace ones that must be retired before they get overused (i.e., for assessment integrity reasons). ■ Create additional sets of assessment videos to use in periodic checks on observer accuracy (e.g., every year or few months, not just at the end of initial observer training). ■ Pre-score videos to use as practice assessments to give observers more opportunities to rate multiple components before their assessment at the end of training. These may include medium-length videos (e.g., 15 minutes).

TIPS

- Pre-scoring a few examples of high-level performance early on in implementation can help raise observers' expectations during training, especially if the examples are drawn from the local context.
- A single video clip can be used to exemplify practice for several components of a rubric. But segments from assessment videos should not also be used as training clips. Observers should not be assessed with video they have already seen.

PUTTING IT INTO PRACTICE

TO LAY THE FOUNDATION:

What parts of your of rubric might you prioritize for pre-scoring to help the most teachers improve their practice?

TO BUILD AND IMPROVE:

What does your observation and training data suggest about prioritizing new videos to better support teachers and evaluators?

*Note: Saving notes in the fields above requires a recent version of **Adobe Reader**.*

How Will You Get the Right Kind of Video to Pre-Score?

A major challenge of pre-scoring is finding enough of the right kind of raw videos of teaching to use in the process. The good news is that changes in technology continue to make it easier and less expensive to capture high-quality video of classroom practice. What once took thousands of dollars of cumbersome equipment can now be accomplished with pocket-sized devices that cost hundreds of dollars—and often with better results. Unfortunately, such advances have yet to produce a plethora of low-cost and easily accessible videos appropriate for pre-scoring. Given that, states and districts will need to invest time in researching possible sources—not just at the outset but on an ongoing basis.

Video for pre-scoring should capture authentic teaching clearly enough to evaluate it.

To understand what makes a video appropriate for pre-scoring, consider how it will be used. Annotated segments will serve as models for how to rate performance in a classroom observation. If rating performance in the classroom depends on how students respond, then how students respond must be clear in the video. If observations are meant to evaluate typical lessons in typical classrooms, then the video used for training should not feature highly unusual lessons or teaching methods. Examples of staged teaching are generally not good for modeling how to rate real lessons that unfold in unpredictable ways. Video for pre-scoring should capture authentic teaching clearly enough to evaluate it.

Don't use video without obtaining consent from all who appear in it. Pre-scored segments that feature real teachers and students are used to illustrate particular performance ratings. Those segments should only be shared in training and with the disclaimer that the ratings only apply to what is featured—not to a teacher's overall ability or to the overall lesson it comes from. But it's still easy to see the breach of trust—and legal issues—that can result if videos of teachers who agreed to be recorded are used in ways for which they did not give permission. Likewise, parents may balk—and file lawsuits—if video that features their children is used in training without their consent. (Ask your system's lawyers to review consent forms before using them.)

A state or district needs to make sure the right written consent is granted, whether it employs its own previously recorded videos, records new lessons for pre-scoring, or purchases access to videos from a fee-based provider. Lesson videos recorded as part of the MET Extension project are available from two such providers: MyLearningPlan, in partnership with ETS, and the SOE Teaching and Learning Exploratory, a platform created by the University of Michigan School of Education. When purchasing access to such videos, check the terms. It's of little use to pre-score video if you can't share it with numerous evaluators in training or if you lose access to the video when it still has value in training.

HOW MUCH RAW VIDEO IS NEEDED?

It's very hard to know ahead of time how many lesson videos will be needed to produce a particular set of pre-scored segments. Often a raw video will include clear-cut examples of just one or two teaching components, but more is possible. (Those with most or all teaching components are best designated as assessment videos.) The yield rate depends largely on video quality and the nature of the teaching captured. A video in which 10 minutes is spent handing out papers might yield an example of low performance on maximizing instructional time but not much else. To find examples of teaching components that show up less often than others also requires reviewing more videos. As a state or district gains experience with pre-scoring, it will learn roughly how many usable segments it tends to get from most lesson videos. This number will likely increase with experience, as school systems get better at identifying raw video appropriate for pre-scoring.

Whether a state or district buys, creates, or repurposes video will depend in large part on available resources. Low-cost options may be needed when starting out and before stakeholders have seen the full benefit of pre-scoring video; the focus at that point is to build an understanding of the process while pre-scoring a small number of videos. School systems may also achieve economies of scale by sharing the products of their pre-scoring, if they use the same rubric. Over time, a school system may tap multiple sources as it learns more about what it needs to enhance observer training. With experience, it also becomes easier to determine which sources are the most dependable in providing video appropriate for pre-scoring.

SNAPSHOT
FILMING CASE STUDIES IN COLORADO

As in other places, Colorado's state evaluation system calls for multiple observations of a teacher over the course of a year. But so far, the state's online platform for sharing pre-scored video has only featured teachers recorded at one point in time. To change that, the state department of education has begun filming in the classrooms of teacher volunteers at different points during the school year. Segments from the recorded lessons will be assembled into video case studies that are pre-scored by state-trained master coders. This will allow users of the state's online evaluator-training platform to practice rating instruction based on different observations at different times in the same classroom. Disclaimers will stress that the segments are excerpted to illustrate specific levels of performance and do not represent overall proficiency of the teacher volunteers.

TECHNIQUES
EXPLORING OPTIONS FOR SOURCING VIDEO

To Lay the Foundation

- Review existing video from local sources, if any exist, to see whether it is of sufficient quality, whether it shows the right content, and whether the permissions granted by teachers and parents would allow its use. (It's possible, but difficult, to get new consents on previously recorded video.)
- Consider to what extent local teachers would consent to have their lessons recorded for pre-scoring. (This might require purchasing video equipment or contracting with a videographer.) Make every effort to ensure a positive experience for such teachers so others step forward.
- Look for opportunities to share video across school systems—either by sharing locally recorded lessons or by sharing in the cost of purchasing access to video from a commercial vendor, such as MyLearningPlan or Teachscape.
- See whether fee-based providers can provide access to a limited number of videos for a minimal cost to build an understanding of the pre-scoring process while producing a starter set to begin training observers.
- Ask school systems with experience pre-scoring what sources have been most useful and cost-efficient for them.

To Build and Improve

- Determine which of the sources you have used is the most consistent in yielding high-quality segments that add value to your observer training. Continue or expand your use of those sources and consider discontinuing the use of ones that less frequently yield good segments.
- Consider how a mix of sources can better enhance training while building out a video library. It's important for evaluators to learn to rate performance with examples from classrooms that look like those in which they will observe. But using examples of low performance from outside the local context is one way to address teachers' concerns about possibly not looking their best in a segment used in their own district.
- Based on your experience pre-scoring lessons from different grades and subjects, consider whether some subjects are better recorded using different techniques (e.g., is a "roving" camera better for science labs or for small-group guided reading instruction?).
- Determine what grades and subjects are least represented (or not represented at all) in current sources. You can ask school systems with more experience pre-scoring what solutions they have found to similar challenges.

PUTTING IT INTO PRACTICE

TO LAY THE FOUNDATION:

What are the most promising sources of video that you might explore for pre-scoring?

TO BUILD AND IMPROVE:

Which sources have been most productive, and how much would additional ones address your most pressing needs?

*Note: Saving notes in the fields above requires a recent version of **Adobe Reader**.*

How Will You Pre-Screen Video for Quality and Content?

Whatever the source of video, pre-screen it. Before considering a video appropriate for pre-scoring, you want to know that someone watching it could determine the performance level based on its content. Assigning someone to make this determination ahead of time saves master coders from struggling with and ultimately rejecting "unscoreable" segments. In a state or school district, where coding is typically done by full-time teachers and administrators, master coders' time is highly limited; to maximize that time, ask others to "separate the wheat from the chaff." Pre-screeners could be district administrators who know the rubric.

One of the best ways to pre-screen a video is to try to rate it; if an early viewer can't see or hear enough to make a judgment, then neither will the master coders. Pre-screening also can be used to identify the most relevant parts of a video. Master coders can then skip over periods when little is happening, such as during announcements or when students are working individually for extended periods and their work cannot be discerned. (Augmenting video with artifacts, such as lesson plans and student work, can help when video alone isn't capturing critical evidence.) Pre-screeners can also prioritize segments that appear to include high-priority examples needed to build out a video library or those that appear to include examples of multiple teaching components and so may be candidates for an assessment video.

Even with pre-screening, master coders will find that some of the segments they are assigned are problematic. They may realize that the audio is not clear at key points or that important pieces of evidence cannot be discerned. This is especially likely when a pre-scoring process is new. Give master coders opportunities to comment on the judgments of pre-screeners and to explain why they think a segment may not be appropriate for pre-scoring. This feedback will improve the process and tools you use for pre-screening going forward.

In a state or school district, where coding is typically done by full-time teachers and administrators, master coders' time is highly limited; to maximize that time, ask others to "separate the wheat from the chaff."

TOOL
DCPS VIDEO-QUALITY CHECKLIST

The District of Columbia Public Schools (DCPS) uses a checklist to pre-screen lesson videos for possible pre-scoring. The tool is meant to determine whether the video provides enough evidence to be scored by a master coder (called an "anchor rater" in DCPS). Items ask, for example, whether an observer could identify the lesson objective and make a judgment about student understanding based on what can be seen and heard. Depending on the answers, a video that's not rejected may be designated as ideal or merely acceptable for pre-scoring. The two tiers allow for prioritizing the most scoreable ones and making sure anchor raters know about any issues with the others.

The complete DCPS Video-Quality Checklist can be found at www.metproject.org.

PUTTING IT INTO PRACTICE

TO LAY THE FOUNDATION:

Who could be assigned the task of pre-screening videos before any are assigned to master coders?

TO BUILD AND IMPROVE:

How could your pre-screening process be improved to further maximize master coders' time?

*Note: Saving notes in the fields above requires a recent version of **Adobe Reader**.*

How Will You Share Video and Collect Rating Justifications?

Pre-scoring involves a two-way transfer of information. You need to get videos to master coders, and master coders need a way to submit their evidence, rationales, and performance ratings for each segment they review.

Readily available tools may be used. A state or district might go with an essentially no-cost option by posting password-protected videos to a Vimeo account or Google Drive and by having master coders submit their work in a simple table using a word processor or spreadsheet. Commercially available platforms (such as MyLearningPlan's Elevate) allow master coders to watch video, take notes, and align relevant evidence to each part of a rubric—all in the same system.

Whatever the system, make sure to test it in the context in which it will be used. If master coders will be reviewing videos on school computers, make sure they're not kept from doing so by the kind of blocking software typically installed on such devices. If they will be meeting as a group, make sure the location has sufficient bandwidth to stream videos. It's a waste of time to assign videos for pre-scoring only to find that participants can't review them and that the work must be put on hold until a solution is found.

Make sure to ask master coders how they're experiencing the process and what might improve that experience. Ongoing efforts to refine the sharing of information may include adding new features, systems, and sources of video.

PUTTING IT INTO PRACTICE

TO LAY THE FOUNDATION:

What tools for sharing video and written score rationales make the most sense for your system to use when just starting out?

TO BUILD AND IMPROVE:

What would make the sharing of video and information with master coders more efficient?

*Note: Saving notes in the fields above requires a recent version of **Adobe Reader**.*

How Will You Recruit Master Coders?

When pre-scoring video, it pays to involve the right people. Master coders (sometimes called "anchor raters" or "master scorers") must be willing to set aside their own ideas of effective teaching and rate what they see based solely on the rubric. They must be open to changing their judgments when presented with convincing evidence. But they also need to be active participants; they can't just passively defer to others. In addition, a master coder must have the patience and attention to detail to repeatedly review the same video to accurately record the behaviors most relevant to scoring each teaching component. When master coders don't buy into their charge, the result can be unproductive and awkward.

ESSENTIAL QUALITIES OF MASTER CODERS

- Deep understanding and appreciation of the observation rubric
- Belief in the importance of consistency in evaluation and feedback
- Strong interpersonal skills and openness to others' opinions
- Good attention to detail, ability to focus, and note-taking skills
- Willingness to commit long-term to an ongoing process done outside of normal work hours

Given the learning curve entailed and that new pre-scored video is always needed, master coding should be seen as a long-term commitment. Most master coders are teachers and administrators who pre-score on a part-time basis. A master coder may take 3–4 hours to review a video individually and another hour to take part in reconciliation. The task can't be done in the middle of typical work days. Often pre-scoring begins over the summer and continues throughout the school year. Any stipends paid are modest compared to the time required. The biggest motivation to participate is the chance to engage with colleagues in rigorous professional learning that's focused on teaching.

> Master coders must be open to changing their judgments when presented with convincing evidence. But they also need to be active participants; they can't just passively defer to others.

The overarching question when planning recruitment is "where will we find the individuals with the best understanding and appreciation of our rubric?" When a rubric is new—and there's no cadre of trained observers from which to select the most skilled—the people with the best grasp of the rubric's indicators may be those who took part in its development, selection, or field testing. As implementation progresses, data become available that can help identify the most accurate observers as potential master coders. As emphasized throughout this guide, it helps to start small. A state or district learns a great deal about what makes for a good master coder from its early attempts to pre-score.

Despite careful recruitment, it may become necessary to discontinue working with a master coder. Individuals who consistently dig in their heels and are unable or unwilling to consider alternative views will not be of help in determining what the right performance rating is for an instance of teaching. Keep in mind that it takes time for new master coders to get comfortable with the process; they will need reminders when they go beyond what's in a video segment and the rubric in rating performance. But after some time, it may become clear that some lack the right mindset or temperament. They may feel that their expertise or authority is questioned by the process. Or they may strongly disagree with what's in the rubric.

TIPS

- Recruiting individuals whose instructional expertise is well known builds credibility for the overall observation system.
- Consider recruiting master coders who represent a variety of stakeholders, including teachers and school leaders from different regions and contexts within the school system. Master coders often become advocates for the process, replicating the use of video to build a common understanding of effective teaching in their schools and districts.
- In some cases, long-held views of effective teaching may make it difficult for some experienced educators to see instruction solely through the lens of a rubric that's new to them. Look for individuals who are able to adjust.

TECHNIQUES
RECRUITING MASTER CODERS

To Lay the Foundation	To Build and Improve
■ If observations are already being done, look for those observers who have demonstrated the most skill in applying the rubric correctly. ■ Recruit from among teachers and administrators who helped develop, select, or pilot the rubric.	■ Retain master coders who demonstrate the ability to work with others to produce clear, evidence-based score justifications. Make sure such individuals feel valued for their work. ■ Recruit from among the most accurate observers, based on observer assessment results.

Note: A cadre of master coders should include experts from different subjects, grade bands, and teaching specialties.

PUTTING IT INTO PRACTICE

TO LAY THE FOUNDATION:

Who in your system has the best grasp of your rubric and would build the most credibility for pre-scoring?

TO BUILD AND IMPROVE:

How can you keep your best master coders and find more like them in more places?

*Note: Saving notes in the fields above requires a recent version of **Adobe Reader**.*

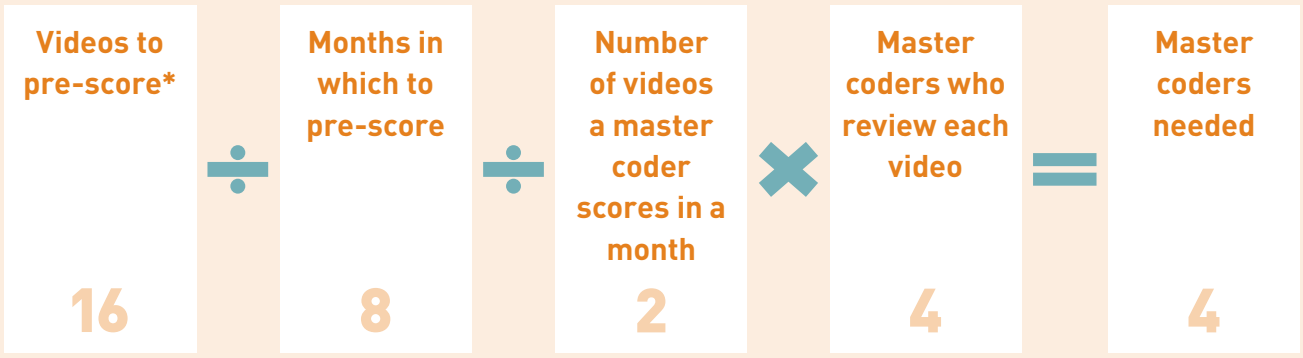
How Many Master Coders Will You Need?

Several factors drive the number of master coders a school system needs. Two of the biggest are the number of videos to pre-score and the amount of time for pre-scoring them. (Another is the number of subjects and grade bands for which evaluators will be trained, as each needs its own set of videos.) Also relevant is how many master coders review each video: Variations of this are possible, but quality assurance requires more than just rating by one pair of coders. (Quality controls are discussed later in this guide.) The example estimate in **Figure 5** assumes that two pairs will review each video before it's used for observer training or assessment.

When determining recruitment needs, consider how much time your master coders will likely be able to devote to the process on an ongoing basis. The estimate shown assumes that a master coder will review two videos each month, which might take a total of 6–8 hours (including reconciliation discussions). If they can review more each month, then they can pre-score more video over the course of the year (or complete the same number of videos sooner). Available resources and current needs will suggest adjustments. The hypothetical school system in the example below might scale up its process the following year by doubling the number of videos to pre-score—resulting in the need for twice the number of master coders. Or a state or district just starting out might only have resources for fewer coders, in which case the number of videos expected to be pre-scored would be reduced.

When starting out, planning is more of a guessing game. You don't know how quickly your coders will be able to work (although they should get quicker with practice). You also don't know how many usable segments of pre-scored video will result from a given set of raw videos of teaching—especially if you're hoping to find specific examples needed in your training. By tracking how long it takes and how much is produced, you'll be better able to plan and improve the process going forward.

FIGURE 5. **FACTORS THAT DETERMINE RECRUITMENT NEEDS**



**Assumes each video is 20–30 minutes. Shorter ones require fewer resources to review.*

PUTTING IT INTO PRACTICE

TO LAY THE FOUNDATION:

What's your estimate for each of the factors that determine the number of master coders needed as you start pre-scoring?

TO BUILD AND IMPROVE:

In what ways could you better estimate recruitment needs or adjust your process to increase productivity while maintaining quality?

*Note: Saving notes in the fields above requires a recent version of **Adobe Reader**.*

How Will You Train Master Coders?

Master coding is a special skill. To be sure, it includes many of the same activities involved in classroom observations carried out for teacher evaluation. Both entail recording evidence and rating practice based on a rubric. But few observers have ever compared notes with others who rated the same teaching and after reviewing it repeatedly on video. What's more, master coding a short segment to show one teaching component at one performance level means setting aside knowledge of what happened at other times in the lesson, whereas in observations, an evaluator considers evidence from the whole lesson. Even the most expert observers face a learning curve when they become master coders.

A major goal of training is to build fluency with the protocol a master coder follows to review video, rate performance, and prepare score justifications for reconciliation. The protocol will vary depending on the extent of pre-screening that a school system does. Master coders can spend more time rating performance if pre-screening identifies which videos may be candidates for assessment videos and which parts of a video can be skipped for lack of relevant content. (See *How Will You Pre-Screen Video for Quality and Content?* on page 17.) Possible assessment videos are those in which an extended segment—perhaps 20 minutes or more—appears to feature examples of most or all of a rubric's components. Regardless, master coders should know what makes for a good assessment video so they can make their own judgments.

TRAINING OF MASTER CODERS SHOULD COVER:

- Using the video-sharing system
- Using common forms/templates to record evidence, timestamps, and score justifications
- Following a protocol for coding
- Writing specific and evidence-based score justifications using the rubric's language
- Reaching agreement on scores and justifications in reconciliation with other coders

The example in **Figure 6** on the next page illustrates a protocol for reviewing a segment pre-identified as a candidate for an assessment video. It begins with identifying instances of specific teaching components and recording detailed evidence with which to determine the performance level of each. Evidence is objective: It might include exact quotes from teachers and students, descriptions of behaviors, or numbers (e.g., how many students responded). Opinions and generalizations, like "the questions were low level" or "the students were on task," won't help evaluators learn to recognize the right indicators of performance—which is fundamental to ensuring consistent ratings and specific feedback.

FIGURE 6. A MASTER CODING PROTOCOL

Steps	Excerpts of What It Might Look Like
1. Review the assigned video, and note when examples of particular teaching components are exhibited.	■ In a 22-minute video, an instance of "checking for student understanding" occurs from 3:23 to 9:57. During that time the teacher circulates among students as they multiply fractions, after which she addresses a student's misconception by explaining another way to solve such problems.
2. For each teaching component noted, record evidence needed to determine the performance level.	■ To one student at each table, the teacher says, "What will you do first?" and "Okay. Why?" ■ At 6:39 the teacher says to the class, "Okay eveyone, I asked Anthony to talk about something he wasn't sure of." ■ At 7:01 the student says, "The answer I got is smaller than the fractions I multiplied. So it didn't seem right." ■ At 7:12 the teacher says, "Okay, Anthony makes a a good point. What's happening when we multiply by a fraction?" ■ At 8:10 the teacher says, "Here's another way to look at it." She draws a 3 X 4 grid on the white board, then says, "Anthony, show us what part of this is 2/3? Great, now what part of that is 3/4? Are we getting smaller or bigger?"
3. Determine the performance level described in the rubric that best matches all the evidence recorded.	■ The evidence aligns with a level 3, of which the rubric says, "the teacher checks for understanding before moving on in a lesson" and "the check for understanding provides information to adjust instruction." ■ It is not a 2, of which the rubric says, "the check for understanding does not provide useful information." ■ Nor is it a 4, of which the rubric says, "the check provided information on most students' understanding." In this segment, the teacher spoke to just one student at each table.
4. Draft a score justification for each teaching component that connects the evidence to the rubric's performance indicators.	■ The teacher's check for understanding provided information to adjust instruction. The teacher checked for understanding by asking one student at each table to solve the problem and explain each step of the solution. ("What will you do first? Okay, why?") At the end the teacher asked a student to explain his confusion to the class ("The answer I got is smaller than the fractions I multipled."), after which she showed a different way to solve such problems (using a rectangular grid).

The second half of the protocol involves determining performance ratings and drafting rationales. For the former, a master coder determines which performance level described in the rubric aligns most with the evidence from the video that's relevant to each teaching component. Rationales for the accurate rating, and for why a higher and/or lower rating would be incorrect, must be clear, concise, grounded in evidence, and written in the language of the rubric. A rationale makes a strong case for a particular rating as representing the correct way to apply a rubric to a specific instance of teaching. It says, in essence: These are the things in the video that make this the right rating, according to this rubric.

Another focus of training is reconciliation. Before a video segment is used to train observers, multiple individuals must agree on the ratings and rationales; a single master coder, no matter how expert, is not enough to ensure accurate ratings for examples to be used across a school system to norm how observers apply a rubric. Variations of reconciliation are possible. Working in pairs—with two master coders comparing notes after rating individually—can be the simplest approach. It also mitigates a tendency among some larger groups to seek consensus rather than make the best judgment based on the evidence. Whatever the structure, guided practice on reaching agreement makes reconciliation run more smoothly.

How a state or district delivers its training to master coders will depend on context and will evolve over time. Geography may limit the extent to which people can be brought together. Webinars and individually completed activities may be used to introduce protocols, video-sharing systems, and for some practice reviewing segments. But a good deal of face-to-face training is recommended when starting out. A state or district is likely to encounter many unforeseen challenges the first

time it trains a group of master coders. It may become clear that the training of master coders needs to better clarify when a particular kind of behavior indicates one teaching component vs. another or the difference between opinion and objective description when justifying scores. Begin with a small group, and learn from an early round of training before formalizing a process to prepare larger numbers of coders.

The exact content of training also will depend on the prerequisite knowledge of participants and how much of the pre-scoring process they'll take on. When a rubric is new to master coders, more time will be needed to review its language before attempting to score. If master coders will recommend whether segments are used as training or assessment videos, then training should cover that process (by determining whether a longer segment includes enough clear examples of different components to serve as an assessment video). If reconciliation will be done without a facilitator—say, with pairs discussing on their own to reach agreement—then training should include practice resolving disagreements and creating coherent justifications.

Regardless of content, facilitated group discussions are especially helpful for training new master coders. Group discussions let them practice justifying their scores with others who have reviewed the same lesson. Pressed by each other, master coders get increasingly more specific in pointing out relevant observed behaviors and in explaining what distinguishes between a rubric's teaching components and performance levels. A good approach is to begin with a facilitator leading whole-group discussion of the same videos, followed by master coders working in smaller groups that more resemble how they'll score video after their training (e.g., with coders reconciling their scores and justifications in pairs).

It may become clear that the training of master coders needs to better clarify when a particular kind of behavior indicates one teaching component vs. another or the difference between opinion and objective description when justifying scores.

TECHNIQUES
TRAINING METHODS FOR MASTER CODERS

To Lay the Foundation	To Build and Improve
Train a cadre of new master coders using a boot-camp[1] model in which: ■ As pre-work, master coders score video segments using the video-sharing system, coding protocol, and template; ■ Master coders meet in group sessions to compare scores and score justifications and to reach agreement by re-examining the segment, the recorded evidence, and/or the rubric indicators; and ■ Master coders leave the sessions knowing what to expect going forward (e.g., how video will be assigned, how reconciliation will take place).	Use differentiated training for new and experienced master coders in which: ■ Highly skilled and experienced master coders help train in the boot camps that continue for new master coders (e.g., to role-play the process); ■ Master coders are given examples of quality score justifications from previous rounds of pre-scoring, with written explanations of why they meet quality criteria; and ■ Returning master coders score pre-scored video to make sure they've maintained their accuracy. A refresher is offered for those who need it.

[1] *For a detailed description of master coder boot camps, see the MET project's* **What it Looks Like: Master Coding Videos for Observer Training and Assessment**, *by Catherine McClellan.*

TIPS

- Whoever leads training should be a highly skilled facilitator. Trainers must be able to guide discussion of segments toward agreement based on evidence, without stepping in to decide what the correct rating should be. When trainers take sides, it doesn't help master coders learn how to weigh evidence and determine a rubric's precise meaning.
- Training works best when it includes review of already pre-scored video. When pre-scoring for the first time, this can be accomplished by asking two to three rubric experts (including the training facilitator, if possible) to rate three to four videos that represent a range of classroom contexts. Even if the facilitator doesn't share the pre-determined ratings, pre-scoring makes it easier to guide discussion to the relevant evidence.
- Begin review of video in training with the most clear-cut examples available. These might be of low-inference components (i.e., observers don't need evidence from different points in a segment to interpret a teacher's intent or student's response).
- If prolonged discussion fails to move training participants closer to agreement, a facilitator should move to another video. When this happens, it should be noted whether the disagreement seemed due to video quality, to some distracting behavior in the lesson, or to difficulty interpreting the rubric. Videos deemed problematic should not be assigned for pre-scoring.
- Reconciliation of evidence is needed even if master coders agree on scores but for different reasons. Both the correct score and the correct justification based on the evidence are needed to support observers in understanding how to apply a rubric to different instances of teaching.

TOOL
RIFT MASTER CODING WORKSHEET

The Rhode Island state affiliate of the American Federation of Teachers (AFT) created a simple paper template to guide master coders when pre-scoring video. The tool has space for evidence, timestamps, agreed-upon ratings, and reasons why ratings other than the correct one would be not be appropriate. Teachers and administrators who engage in pre-scoring organized by the union complete the form when reviewing video in three-hour work sessions. Review and reconciliation of ratings takes place in pairs. The Rhode Island Federation of Teachers and Health Professionals (RIFT) started pre-scoring video as part of the AFT's Investing in Innovation (i3) grant from the U.S. Department of Education. New York State United Teachers also participates in the grant and is engaged in pre-scoring.

DCPS ADVICE FOR FACILITATING RECONCILIATION

In the District of Columbia Public Schools (DCPS), reconciliation takes place in facilitated phone calls after score justifications are submitted by each pair of master coders (called "anchor raters" in DCPS). Facilitators are central office administrators who lead implementation of the district's classroom observation system. To assist them in bringing raters to agreement, the district developed a two-page set of facilitation tips. Among the advice: Begin each session with a reminder of goals and norms (e.g., "be willing to consider evidence from a different perspective"); start with areas of agreement; and summarize apparent disagreements in evidence or interpretation before asking raters to clarify them.

DCPS GUIDELINES FOR SCORE RATIONALES

DCPS provides its anchor raters with a two-page set of guidelines for drafting score justifications. Included are annotated examples of justifications that meet specific criteria, such as beginning with the language of the rubric when describing evidence and including any specific evidence relevant to determining the performance level. Examples of justifications that don't meet the criteria also are provided.

All these tools may be found at **www.metproject.org**.

PUTTING IT INTO PRACTICE

TO LAY THE FOUNDATION:

Who has deep knowledge of your rubric, strong facilitation skills, and might be able to lead training of master coders?

TO BUILD AND IMPROVE:

Where are the biggest gaps in the knowledge and skills of your master coders, and how might you fill them?

*Note: Saving notes in the fields above requires a recent version of **Adobe Reader**.*

How Will You Check the Work Of Master Coders for Quality?

When quality is important, quality controls are essential. No matter how carefully a pair of master coders reviews a video and reaches agreement on ratings, additional checks on their work are needed before a segment is used in observer training or assessment. With checks on quality, a state or district has greater confidence that other experts in the rubric could replicate the right ratings. Without them, there's an increased chance of video being used in observer training that confuses evaluators or even promotes incorrect application of a rubric. Quality-control checks also provide more timely information for improving a process than waiting until the use of videos reveals problems.

Quality-control checks also provide more timely information for improving a process than waiting until the use of videos reveals problems.

How a school system checks for quality will depend in part on resources and whether its pre-scoring process is new or well developed. When just starting out, a sufficient number of master coders may not be available to confirm the reconciled ratings for videos that they weren't assigned to pre-score. Instead, experts in the rubric may be assigned to review the evidence and rationales agreed upon in reconciliation to make sure they align with the rubric (i.e., do the behaviors described match how the rubric defines performance for the rating given?). Later, a school system can assign pre-scored segments to additional master coders, or to trained observers, to see whether they produce the same scores.

TECHNIQUES ENSURING QUALITY

To Lay the Foundation	To Build and Improve
■ Remove videos that master coders struggle to score. (But first determine whether the problem is a lack of clarity in part of the rubric or misinterpretation by one master coder.)	■ Continue to remove videos from the process that master coders cannot score.
■ Have additional experts in the rubric review all score justifications to make sure that they align with the rubric criteria (i.e., does the evidence cited align to the rubric's criteria for the score given?).	■ Before using videos in training, have them scored by additional master coders or observers to see whether they assign the same scores.
	■ Assign someone to review existing pre-scored video for potential problems (e.g., distracting behaviors, outdated references or materials).

TIP

The same master coders should not always take part in reconciliation together. Assignments should regroup master coders periodically to avoid the unconscious tendency to defer based on familiarity.

PUTTING IT INTO PRACTICE

TO LAY THE FOUNDATION:

What tools and procedures could you use for quality control as you start a pre-scoring process?

TO BUILD AND IMPROVE:

How could additional quality controls tighten up your process and build greater confidence in what's produced?

*Note: Saving notes in the fields above requires a recent version of **Adobe Reader**.*

How Will You Keep Track Of the Process and Products?

An ongoing process requires ongoing management. Someone needs to own responsibility for making sure the activities discussed throughout this guide take place, and according to schedule. That means raw video must be acquired and pre-screened before assignments are made to master coders, and master coders must know when to expect new assignments, when their work is due, and when they need to take part in discussion to reconcile disagreements in scores and score justifications. After reconciled ratings are confirmed, someone needs to make sure the reconciled ratings and evidence—along with the video segments they refer to—are organized and accessible for use in observer training and assessment.

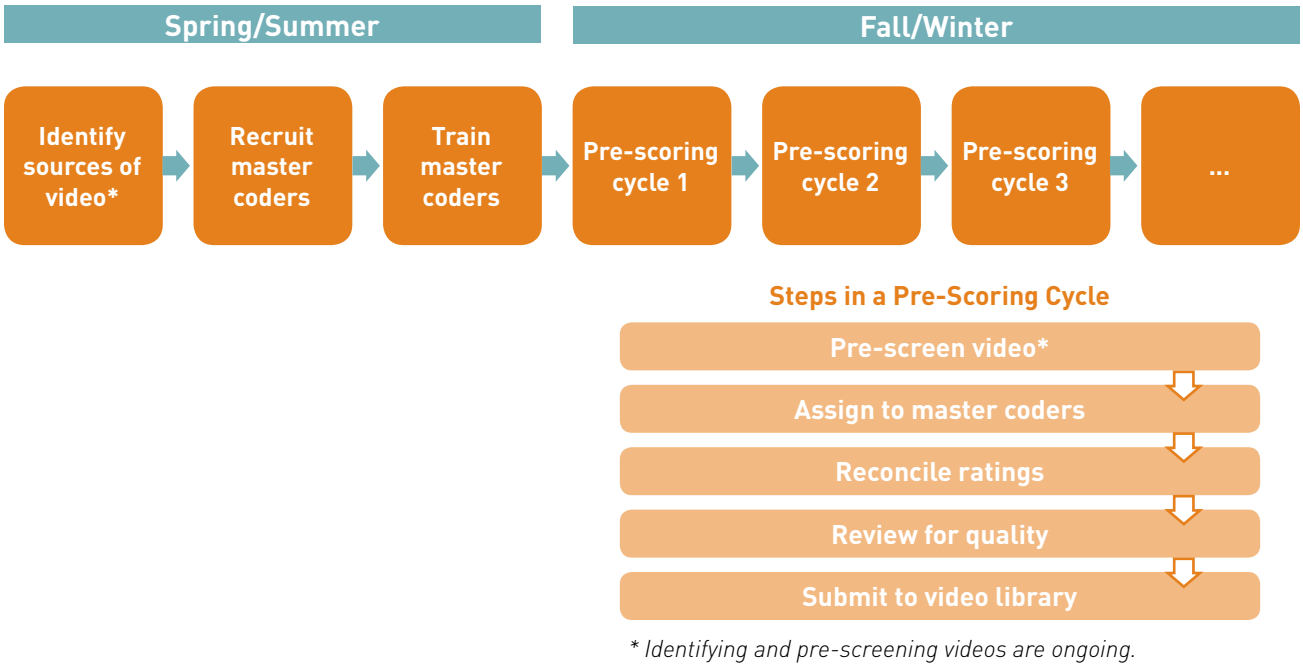
On the next page, **Figure 7** suggests a general schedule of activities throughout the year. Because most master coders are limited in the number of hours they can devote to pre-scoring, pre-scoring typically takes place in a series of cycles, each lasting perhaps several weeks. The number of cycles will vary depending on available resources and the number of videos a school system plans to pre-score in a given year. Some activities may overlap or take place on a rolling basis; pre-screening can be ongoing so long as enough videos are pre-screened to make assignments to master coders at the start of each cycle. Often master coder training takes place over the summer when schedules are more flexible.

Maintaining the catalog of videos and ratings produced is important to a library's development. From the beginning, records should be kept of any segments that master coders were unable to rate; the video should be retired if the culprit is video quality (and not a lack of understanding of the rubric among the master coders). When building out a library, knowing which parts of a rubric are already illustrated with examples—and which grades and subjects are represented—can guide subsequent efforts to fill in the gaps. Keeping track of assessment video usage will help identify which ones need replacing before overexposure results in assessment-integrity issues.

KEY MANAGEMENT TASKS IN PRE-SCORING VIDEO

- Scheduling the process, from pre-screening and the initial training of master coders through each cycle of master coding, reconciliation, and quality checks
- Communicating expectations and deadlines to all participants
- Assigning video to master coders
- Organizing finalized ratings, justifications, and videos for use in observer training and assessment
- Keeping track of who played what role in pre-scoring each video segment, and gathering information about each segment's use in training
- Removing problematic videos from the library, and identifying assessment videos for replacement

FIGURE 7. **AN ONGOING PRE-SCORING PROCESS**



** Identifying and pre-screening videos are ongoing.*

PUTTING IT INTO PRACTICE

TO LAY THE FOUNDATION:

Who needs to take part, and when, in the key steps involved in pre-scoring over the course of the year?

TO BUILD AND IMPROVE:

How could specific steps of the pre-scoring process over the year be managed differently to improve quality and efficiency?

*Note: Saving notes in the fields above requires a recent version of **Adobe Reader**.*

How Will You Use Data for Continual Improvement?

Information makes improvement possible. As emphasized repeatedly in this guide, a state or district's capacity to pre-score video is built through continual improvement. Indeed, the same goes for the capacity to deliver training that makes use of pre-scored video, for the capacity to ensure that evaluators are effectively coaching teachers, and for every other capacity required in an observation system that aspires to support great teaching. But whether each attempt to pre-score video is better than the last hinges on collecting the right information at the right time, asking the right questions, and making changes when warranted.

> Plans should be in place from the outset of the first round of pre-scoring to track information on videos, on participants in the pre-scoring process, and on the videos' subsequent use in training and assessment.

Collect data from the beginning. Plans should be in place from the outset of the first round of pre-scoring to track information on videos, on participants in the pre-scoring process, and on the videos' subsequent use in training and assessment. Otherwise, there's no way to know what worked and where something different is needed. You can't leverage the most successful pre-screening practices, or the most successful pre-screeners, if you don't keep track of who pre-screened what and how often they succeeded in identifying useful segments. You can't identify the most skilled master coders if you don't keep records of which master coders submitted what ratings for reconciliation.

The table on the following page lists the key information to collect from the start of a pre-scoring process and the key questions to consider at each subsequent stage of implementation. Some of the latter are variations of the prompts included throughout this guide under "To Build and Improve." The main unit of analysis is the video segment. For each completed pre-scored segment, a state or district should be able to call up where it came from, who played what part in pre-scoring it, and how well it served its purpose. As noted, the data include quantitative, categorical, and anecdotal information. As with teaching and learning, improvement of a pre-scoring process requires multiple types of evidence.

TECHNIQUES
MANAGING INFORMATION FOR CONTINUAL IMPROVEMENT

To Lay the Foundation

Keep track of the following for each video segment.

From the pre-scoring process:

- **Segment name.** e.g., "6th Grade Pollination Lesson: Use of Questioning"
- **Possible use.** Whether as a training video to show one component at one performance level, as an assessment video to practice rating multiple components, or other
- **Start and stop times.** When in the lesson video the segment begins and ends, e.g., "14:02–17:49"
- **Grade, subject, and topic.** e.g., "6th Grade, Science, Pollination"
- **Teacher and school.** e.g., "Jane Apple, ABC Middle School"
- **Lesson video.** A name for the longer unscored video from which the segment is excerpted, e.g., "6th Grade Pollination Lesson"
- **Video source.** Where the video came from, e.g., MyLearningPlan, recorded by district, or some other
- **Pre-screener.** Who reviewed the raw video ahead of time for quality and, if part of the pre-screening process, pre-determined segments for pre-scoring
- **Master coders and ratings.** Names, plus ratings provided by each for each component of teaching
- **Reconciliation results.** Date of discussion and agreed-upon ratings.
- **Quality reviewers.** Who did what, and when, to check the work of master coders, e.g., "Use of rubric terms checked by John Doe; backscoring performed by Julie Smith."

From use in training/assessment:

- **Trainees' scores.** Scores given by observers in training/assessment for each teaching component and date given
- **Challenges noted.** e.g., "Trainees could not hear enough table discussion to collect evidence on checking for student understanding" or "Score justifications from master coders did not cite sufficient evidence for trainees to understand rationale."
- **Strengths noted.** e.g., "Clear evidence to help trainees see the difference between questions to push student thinking and questions to check their understanding."
- **Recommendation for continued use.** For same use in training, for different use, or for discontinued use; e.g., "Evidence of both high and low performance indicators makes this better for after trainees have been exposed to more clear-cut examples."

To Build and Improve

Consider the following questions to plan improvements.

- **Which segments worked best and should be reused? Which should be "retired" or used for a different purpose?** Look for segments that trainees found especially helpful in understanding the rubric and distinctions between score points. Conversely, if many trainees could not reproduce the correct ratings for a segment, review and consider whether the challenge relates to video quality, the presence of conflicting evidence, or clarity of the ratings and justifications produced by the master coders.
- **Which video sources are the most productive?** Determine whether some sources yield more usable and helpful segments than others.
- **Which pre-screeners are the most skilled?** Did the segments that were most successful in training come from particular individuals? If so, consider whether they can do more of the pre-screening and whether they are using techniques that others could use.
- **Which master coders were the most successful?** See which master coders most often submitted ratings that were the same as those determined after reconciliation and quality reviews. See whether some master coders more frequently submitted rationales that were clear, concise, and grounded in evidence and the rubric's terms.
- **Were some teaching components harder to rate correctly than others?** For those that were repeatedly challenging to participants in training, consider whether additional non-video evidence is required (e.g., lesson plans) or whether the footage needs to be captured differently (e.g., with student interviews about what they're learning).
- **Were some grades and subjects harder to score correctly than others?** Look for trends among the least successful segments. Consider whether changes in the nature of the video used in training might address the issue (e.g., with a roving camera operator).

Continue to collect the information listed in the column to the left.

PUTTING IT INTO PRACTICE

TO LAY THE FOUNDATION:

How could you begin to collect key information to evaluate your pre-scoring process?

TO BUILD AND IMPROVE:

What additional information-gathering and analysis would better support your system's continued improvement?

*Note: Saving notes in the fields above requires a recent version of **Adobe Reader**.*

Concluding Thoughts: Learn through Experience

Knowledge sharing is essential to quality implementation of any new practice, and among promising practices to improve teaching and learning, pre-scoring video is relatively young. This guide shares the knowledge of experts involved in pre-scoring for research, assessment, and professional development. But in no way could these pages cover all there is to know or all that might be helpful. Three questions guided what to include:

- What do practitioners most need to know to get started?
- Where are the biggest pitfalls, and how can they be avoided?
- How can practitioners set themselves up for continual improvement?

The rest must come by experience. Most of a school system's knowledge about how to pre-score video will be the result of learning by doing. Each activity undertaken as part of the process is an opportunity to better understand what works and where something different is needed. States and districts will encounter problems not addressed in this guide, and the right solutions will vary by context. Networking with others implementing the process will accelerate this learning.

Pre-scoring video is complex and challenging. But while it takes time to master, the process often hooks participants right from the start. That's because pre-scoring offers the rich opportunity to engage with other professionals in a rigorous analysis of teaching and learning. The great value in sharpening a shared vision of effective teaching becomes quickly apparent.

Most of a school system's knowledge about how to pre-score video will be the result of learning by doing. Each activity undertaken as part of the process is an opportunity to better understand what works and where something different is needed.

Planning Worksheet

Create a plan to start or improve a pre-scoring process by filling in the spaces below with key decisions and dates, action items, responsible parties, next steps, etc. Refer to the pages cited for practical insights, techniques, models, and tips. The page number next to each question links to the appropriate guide section. To return to this worksheet, click on "Making It Real" at the bottom of any page.

GETTING READY	
p. 5	**How will you create support for pre-scoring video?**
p. 8	**What is your goal for pre-scoring video this year?**
p. 11	**What pre-scored video do you need?**
GETTING VIDEO	
p. 14	**How will you get the right kind of video to pre-score?**
p. 17	**How will you pre-screen video for quality and content?**
p. 19	**How will you share video and collect rating justifications?**

*Note: Saving notes in the fields above requires a recent version of **Adobe Reader**.*

PREPARING PARTICIPANTS	
p. 20	How will you recruit master coders?
p. 22	How many master coders will you need?
p. 24	How will you train master coders?
GETTING BETTER	
p. 29	How will you check the work of master coders for quality?
p. 31	How will you keep track of the process and products?
p. 33	How will you use data for continual improvement?

*Note: Saving notes in the fields above requires a recent version of **Adobe Reader**.*

ABOUT THE MET PROJECT

The MET project was launched in 2009 as a research partnership of academics, teachers, and education organizations committed to investigating better ways to identify and develop effective teaching. Culminating findings from the project's three-year study were released in 2012. Funding came from the Bill & Melinda Gates Foundation. For more, see **www.metproject.org**.

ABOUT THE AUTHORS

- **Jeff Archer** is an independent communications consultant who specializes in school improvement issues. A former journalist, he has worked extensively with the MET project to help translate the implications of its research for practitioners.
- **Steve Cantrell,** a former middle school teacher, is a senior program officer in the College Ready division at the Bill & Melinda Gates Foundation. He has played a leading role in guiding the foundation's investments in innovative research and development focused on improving teaching and learning. He co-directed the MET project's groundbreaking study of teacher evaluation methods.
- **Steven L. Holtzman** is a senior research data analyst in the Data Analysis and Research Technologies group at Educational Testing Service (ETS). He served as a research lead on the MET project's video-scoring project and several other MET project studies.
- **Jilliam N. Joe** runs a research and assessment consulting business, Measure by Design. She is a former research scientist in the Teaching, Learning, and Cognitive Sciences group at ETS. At ETS, she provided psychometric and design support for the MET video-scoring project and research and implementation direction for the Teachscape Focus observer training and certification system.
- **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 system.
- **Jess Wood,** a former middle school teacher, led development of an online observer training system while at the District of Columbia Public Schools. 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.

Some material in this guide is borrowed from the MET project brief ***What It Looks Like: Master Coding Video for Observer Training and Assessment***, by Catherine McClellan of Clowder Consulting. In her former role at ETS, McClellan led the team that designed the online observer training, assessment, and scoring system used by ETS for the MET project's classroom observation research. She provided valuable feedback on an early version of this guide.

Acknowledgments: The following people reviewed drafts of this guide and gave suggestions that led to innumerable improvements: Stephanie Aberger (District of Columbia Public Schools); Katy Anthes and Dawn Pare (Colorado Department of Education); Sara Heyburn (Tennessee Department of Education); Dawn Krusemark (American Federation of Teachers); Angela Minnici (American Institutes for Research); David Steele (retired from Hillsborough County Public Schools, FL); and Jonathan Stewart (Partnerships to Uplift Communities [PUC] Schools). Independent consultant Mark Toner helped immensely with editorial guidance and thought-partnering. Significant editorial and design assistance came from KSA-Plus Communications.

BILL & MELINDA GATES *foundation*

Guided by the belief that every life has equal value, the Bill & Melinda Gates Foundation works to help all people lead healthy, productive lives. In developing countries, it focuses on improving people's health and giving them the chance to lift themselves out of hunger and extreme poverty. In the United States, it seeks to ensure that all people—especially those with the fewest resources—have access to the opportunities they need to succeed in school and life. Based in Seattle, Washington, the foundation is led by CEO Susan Desmond-Hellmann and Co-chair William H. Gates Sr., under the direction of Bill and Melinda Gates and Warren Buffett.

November 2014

BILL & MELINDA GATES *foundation*

www.gatesfoundation.org