MET project

POLICY AND PRACTICE BRIEF

BILL & MELINDA GATES foundation

Foundations *of* Observation

Considerations for Developing a Classroom Observation System That Helps Districts Achieve Consistent and Accurate Scores

Jilliam N. Joe | Cynthia M. Tocci | Steven L. Holtzman | Jean C. Williams

ETS, PRINCETON, NEW JERSEY

ABOUT THIS DOCUMENT: This paper, written by members of the ETS team that built the MET project's observation scoring system, offers guidance to school system leaders on the elements of observer training and assessment that can produce accurate and reliable results for teachers. Briefs and research reports on the MET project's analyses of classroom observation instruments, and of other measures of teaching, may be found at www.metproject.org.

ABOUT THE MET PROJECT: The MET project was a research partnership of academics, teachers, and education organizations committed to investigating better ways to identify and develop effective teaching. Funding was provided by the Bill & Melinda Gates Foundation.

The approximately 3,000 MET project teachers who volunteered to open up their classrooms for this work were from the following districts: The Charlotte-Mecklenburg Schools, the Dallas Independent Schools, the Denver Public Schools, the Hillsborough County Public Schools, the Memphis Public Schools, the New York City Schools, and the Pittsburgh Public Schools.

Partners included representatives of the following institutions and organizations: American Institutes for Research, Cambridge Education, University of Chicago, The Danielson Group, Dartmouth College, Educational Testing Service, Empirical Education, Harvard University, National Board for Professional Teaching Standards, National Math and Science Initiative, New Teacher Center, University of Michigan, RAND, Rutgers University, University of Southern California, Stanford University, Teachscape, University of Texas, University of Virginia, University of Washington, and Westat.

ACKNOWLEDGMENTS: The authors would like to thank Courtney Bell, Teresa Egan, Don Powers, Richard Tannenbaum, Caroline Wylie, and June Zumoff for their thoughtful feedback during the preparation of this manuscript. We would also like to thank Marie Collins for her invaluable contributions to the editing process.

June 2013

Table of Contents

Introduction

The purpose of this paper is to provide states and school districts with processes they can use to help ensure high-quality data collection during teacher observations. ETS's goal in writing it is to share the knowledge and expertise we gained (a) from designing and implementing scoring processes for the Measures of Effective Teaching (MET) project, which used content-neutral and content-specific observation instruments for measuring teacher practice, and (b) from other experiences we have had with observation systems. We broadly outline what is involved in setting up such a system, including considerations related to selecting an observation instrument, training observers, certifying observer proficiency, and conducting post-training observations of classroom practice. This paper does not include discussion of how to give effective feedback based on observation data. Providing feedback to teachers was not within the scope of the MET project and, thus, it is not a component of this document. Feedback training is an important part of the observer training continuum and should be included in your observation system.

Until recently, observation practices often drew on as many different conceptions of "good teaching" as a district had observers. By contrast, a well-selected observation instrument provides a shared conception of teacher effectiveness that standardizes the lens through which observers view teaching practice; provides teachers with meaningful data to improve their practice; and has the potential, ultimately, to help improve student learning.

As we stress in the pages that follow, the first step is to employ a research-based observation instrument, as it is the core of high-quality teacher observations. Until recently, observation practices often drew on as many different conceptions of "good teaching" as a district had observers. By contrast, a well-selected observation instrument provides a shared conception of teacher effectiveness that standardizes the lens through which observers view teaching practice; provides teachers with meaningful data to improve their practice; and has the potential, ultimately, to help improve student learning.

Two additional steps we discuss are also critical to developing a valid and reliable observation system. The first, training observers, helps ensure that everyone has the same understanding of teacher quality for observation purposes. The second, verifying that observers are calibrated to the instrument's scoring levels, helps ensure the accuracy and reproducibility of the observation data that your system collects.

We conclude the paper with an examination of the issues that can emerge after an observation system is in place and what a district can do to support and monitor observers on an ongoing basis. As the stakes assigned to observation data increase—and they will increase—these issues have the potential to become more litigation-prone. We give districts some food for thought that may help them build robust observation systems.

In offering what we see as "best practices," it is important to clarify—given the limited research base that exists to substantiate the processes and recommendations we set forth—that these are *not* strict guidelines. Our intent, as the saying goes, is not to "make the perfect the enemy of the good." Rather, the processes we propose are designed to help districts move from a casual system of observation toward more rigor, standardization, and objectivity. The three basic measurement principles of assessment—reliability, validity, and fairness—underlie the processes we suggest. However, in applying these principles, it is important to understand that they serve to control, rather than eliminate, the inherent subjectivity of the observation process. You should consider the development of your district's observation instrument and procedures as an iterative process—one continually subject to refinement and calibration. As you gather data from observer training, the certification process, and live observations you will be able to make more informed decisions about any changes that might be necessary.

The Observation Instrument

THE FOUNDATION OF A RELIABLE AND VALID OBSERVATION SYSTEM

As noted in the introduction, a well-developed and -implemented teacher observation instrument can (a) standardize the lens through which observers view teaching practice, (b) provide teachers with meaningful data to improve their practices, and (c) ultimately strengthen student learning outcomes. Investing thought in the design or selection of an observation instrument and its implementation procedures can help ensure that the inferences your district makes about the quality of your teachers' practice are reliable and valid. In this section, we offer general guidelines and principles for the development and selection of an observation instrument and provide information about the observation instruments used in the MET project.

OBSERVATION IN A TEACHER EVALUATION SYSTEM

Before we begin, let us unpack what we mean by "teacher observation instrument." It is worth noting that teacher evaluation systems tend to have three major contributors: (1) classroom observation, (2) other measures of professional practice, and (3) student outcomes/achievement. The quality of each of these sets of data influences the validity of teachers' overall effectiveness evaluation.

Two kinds of performance assessment instruments can reside within a teacher evaluation system: (a) the classroom observation instrument and (b) other measures of professional practice. The purpose of the observation instrument is to measure *observable* behaviors of teaching practice and classroom interactions—

MET Project Example

Quality Classroom Observation Instruments

One purpose of the MET project, in addition to determining which observation measures best identify what teachers do that helps students learn, was to add to the research base for the instruments used in the study. The instruments used were the Framework for Teaching (or FFT, developed by education consultant Charlotte Danielson), the Classroom Assessment Scoring System (or CLASS, developed at the University of Virginia by Robert Pianta, Karen La Paro, and Bridget Hamre), the Protocol for Language Arts Teaching Observations (or PLATO, developed by Pam Grossman at Stanford University), Mathematical Quality of Instruction (or MQI, developed by Heather Hill at Harvard University), and Quality Science Teaching (or QST, developed by Raymond Pecheone and Susan E. Schultz at Stanford).

Each instrument used was carefully molded for the MET project. Dimensions of practice were eliminated from the instruments used in the study if they (a) could not be assessed by raters with no prior knowledge of the teacher or contextual information about the students in the class, (b) required raters also to examine artifacts such as lesson plans, or (c) required raters to make inferences about the teacher that were not informed by evidence collected while watching 15 to 30 minutes of video.

For example, Charlotte Danielson's FFT instrument (Danielson, 2010) underwent careful scrutiny by Danielson and other FFT experts to identify those dimensions of practice that, while suitable for an overall evaluation system, were not suitable for observation. Consequently, Domains 1 (Planning and Preparation) and 4 (Professional Responsibilities), which required teacher lesson plans as a primary source of evidence, were omitted from the observation instrument. Other changes in Domains 2 and 3 were made as well. For example, the component about physical space was not included and the components about classroom assessment and responsiveness were combined.

that is, to quantify the quality of teaching practice. The function of the other measures of professional practice, on the other hand, is to codify and measure activities related to teaching practice that are not observed in the classroom. The focus of this paper is the classroom observation instrument.

The two instrument types specific to the direct measurement of teaching practice are further delineated as follows:

- Teacher observation instruments measure observable *behaviors* of teaching practice that are demonstrated in the classroom, such as student engagement, use of questioning techniques, classroom management, and accuracy of presentation of content. It bears mentioning that student perception surveys are a class of teacher observation instruments. They, too, can be useful indicators of teaching effectiveness when combined with other measures (Bill & Melinda Gates Foundation, 2012). It is not within the scope of this paper to discuss the development of those instruments, however.
- Other instruments or surveys measure performance indicators or outputs, such as lesson planning, engagement with community, facilitation of parent-teacher meetings, professional development courses completed, and leadership roles assumed over the course of the year.

DEVELOPING A TEACHER OBSERVATION INSTRUMENT

Districts and other local entities (such as consortia of districts, counties, and states) that wish to use an observation instrument can choose between three approaches. The first approach is to develop an instrument from the ground up. The second is to select a publicly available instrument that closely aligns with and captures district or state standards for teaching practice. The third approach, which is common to the field, is to modify an existing instrument to achieve alignment between it and district or state standards for teaching practice.

Developing an observation instrument locally has its advantages. The locally developed instrument allows stakeholders to provide broader input about the characteristics of and expectations for teaching practice. Engaging together in this process could help promote desired professional growth in your staff with respect to teaching practice. Inviting administrators, school leaders, and teachers to participate in the development process could help ensure that the instrument reflects district or state standards for teaching practice, which in turn can garner buy-in and support for the use of the instrument in the evaluation process. Depending on whether the instrument is being developed at the district or state level, the development process may also serve as a catalyst for aligning district and state standards to the Common Core State Standards.

Despite all of these positive outcomes, local instrument development also poses a number of challenges. While the locally developed instrument can help with consensus-building, if it is not aligned with adopted statewide standards (measured in the same way and along the same scale), opportunities for valid cross-district or other within-state comparisons of teaching practice will be limited. Also, while building an observation instrument from scratch requires a significant initial investment of time, money, human capital, and other costs, the resource implications do not end there. Given the increasing stakes associated with teacher observation data, the completed instrument must be subjected to study to generate evidence that supports the validity of the decisions that are intended to be made based on scores from that instrument. Validity is not an intrinsic property of an observation instrument or score; it is a contextual property that also depends on how the scores are used (McClellan, Atkinson, & Danielson, 2012). The "validity argument" is strengthened through

an ongoing and systematic collection of evidence that supports the interpretations and decisions that are made based on those scores (Kane, 2001). If the aspects of teaching practice that are important to student learning outcomes change and the instruments do not reflect that change, validity of score interpretations will be affected. Professional development, emerging theories, and federal law can change the nature of what we think of as effective teaching so much so that the instrument no longer captures the full range or depth of teaching practice. (See Bell, et al., 2012; Kane, 2001, 2006; and Messick, 1995, for more comprehensive discussions of validity.)

When considering developing an observation instrument locally, it is wise to have a plan in place for studying the reliability and validity of any results that stem from the use of your observation instrument. Here are some aspects of validity you will need to consider as you design a data collection and research plan:

- **Content-based validity:** Evidence that confirms that the observation instrument measures what it was designed to measure and is representative of the content area. A measure of observed teaching practice would have reasonable support for content-based validity if it included dimensions of professional practice that are theoretically and empirically reflective of teaching practice in the classroom. Content-based validity would be compromised if your observation instrument was designed to measure teacher quality, but it was also detecting a curriculum effect (e.g., teachers with well-aligned curricula score high and those with poorly aligned curricula score low).

- **Structural validity:** Evidence that confirms that the dimensions of the instrument relate to one another in expected ways. When observation data are analyzed, you should find that they are structured in a way that represents the theoretical or expected relationships among the dimensions of practice. Inter-item correlations and exploratory and confirmatory factor analyses are often used to address this aspect of validity.

- **Convergent validity:** Evidence that confirms that the dimensions of the observation instrument are correlated with other measures of classroom teaching practice in expected ways (e.g., the relationship between observation and student perception survey data). This aspect of validity supports interpretations that can be made based on observation data.

- **Criterion-related validity:** Evidence that confirms that the dimensions of the instrument are related to other criteria for teaching performance (provided these criteria are sufficiently reliable), such as student perception surveys and valued-added measures.

- **Generalizability:** Evidence that confirms or supports the generalizability of interpretations made based on the scores from a sample of all possible dimensions, observation occasions, lessons, and observers. Sampling can lead to error, and this error can weaken the validity argument for the score interpretation. Generalizability theory is commonly used as a framework to examine the impact of measurement error that stems from sampling from the universe of all possible measurement conditions. Here are possible sources of error in the observation system:

  Dimensions. It would be unreasonable to expect an observer to assess all possible dimensions of teaching practice with one instrument. Therefore, the dimensions of teaching practice generally found in an observation instrument are thought to be a sample of all possible dimensions in the target domain.

Lessons. It would also be unreasonable to assess all lessons, or perhaps even the entirety of a lesson, for a given teacher. As such, we often must choose a sample of all possible lessons or a sample of possible time points within a lesson from which to measure and make interpretations about teaching practice.

Observers. In many instances in practice, it would also be unreasonable to have all possible trained observers observe a single teacher. Therefore, teachers should be observed by enough individuals to sufficiently reduce any variance in scores due to who does the observing.

Guidelines for Instrument Development

Unless your district is prepared to invest the level of resources required to develop a sound observation instrument, we strongly encourage you to seek an available instrument that closely aligns with your district's standards for teaching practice. However, if developing an observation instrument is the direction your district decides to take, we suggest that you start by doing the following:

- Define and document the dimensions of teaching practice and level of content-specificity the instrument should capture, as well as the general constraints under which those behaviors will be observed (e.g., length of the observation). This document will provide you with a general structure you will need.
- Consider how the observation data will be used. For example, if the observation data are intended to support feedback discussions, then the criteria for judgment will need to be thorough enough to enable the principal or instructional leader to direct the teacher to specific behaviors that characterize performance along the continuum (e.g., weak, basic, exceptional).
- Once the dimensions are defined, determine the scoring scale and criteria (the rubric) observers will use to judge the quality of teaching practice along each dimension. The length of the scale should be determined by how many score points are necessary to meaningfully distinguish different levels of practice (Baldwin, Fowles, & Livingston, 2005; Lane & Stone, 2006).
- Gather feedback from subject-matter or content experts to confirm that the instrument and its rubrics appear to measure what they are intended to measure.
- Try applying the instrument and its rubrics to classroom observation to determine what is working, what is not working, and what modifications have to be made. What to investigate during the try-out is discussed in more detail in the "Piloting Your Instrument" section.

Even after the initial development phase, ongoing work is needed to ensure that the instrument remains a useful measure of teaching practice.

SELECTING A PUBLICLY AVAILABLE INSTRUMENT

If you have concluded that developing an observation instrument requires more time and other scarce resources than you have available, then you may wish to consider the alternative approach: selecting a publicly available, research-based classroom observation instrument.

Selecting a publicly available instrument that has been studied, refined, and used in practice has its advantages. Maybe you have begun the instrument development process, but legislative demands and resource needs have impeded the process. Or, like many districts, you may be facing aggressive implementation deadlines for your observation system. Adopting a publicly available instrument for one or two years could give you time to move through at least one development cycle of your own instrument without affecting the quality of observations. It may even be an effective way of educating everyone in the district about what is involved in developing a local instrument before actually undertaking that task. However, moving from one instrument to another is not a trivial step, and there can be several drawbacks. For example, it requires a reinvestment of costs related to observer training and certification. In addition, any major change to the scoring instruments or procedures may result in incomparable pre- and post-change teacher observation data. Ultimately, adopting a publicly availably instrument of course means adopting the lens through which the developers view teaching practice. Depending on how closely this view aligns with district or state standards, training observers to reliably apply that instrument to live observation can be challenging.

A Suggested Selection Plan

When selecting a publicly available teacher observation tool, you should examine the available instruments with the following questions in mind:

- Does the instrument measure the behaviors of classroom teaching practice that are observable? (If not, it should not be considered for use in classroom observation or it should be modified to include only behaviors that are observable.)
- Does this instrument measure the behaviors of classroom teaching practice that are important to student learning outcomes? (If not, it should not be considered for use in classroom observation.)
- Are the included behaviors of teaching practice observable within a typical classroom observation time period? (They must be to be considered for use in classroom observation.)
- Do the criteria for judgment require support materials (e.g., lesson plans, student work, and communication with parents) or prior knowledge of other teacher competencies and behaviors that might be obtained beyond the typical classroom observation period? (If yes, these criteria would be better assessed through an ongoing and more frequent evaluation framework.)
- Is there sufficient validity evidence to support the intended use and interpretation of the observation data? (You can find out by conducting a search for available reliability and validity studies using the instrument.)

After analyzing the instrument options, have a conversation with representatives from districts currently using the instrument you are considering. Try to get a sense of the level of reliability their observers have been able to achieve with the instrument (i.e., from observer to observer). Ask them to describe the conditions under which they gather data. And encourage them to share the challenges and successes they encountered while training observers to use the instrument.

Finally, talk to the instrument developer(s) to get a sense of any modifications planned for future versions of the instrument. Ask them to discuss the motivation for these modifications and how they expect the modifications to improve the quality of the resulting observation data. Perhaps the current version of the instrument does not completely align with your district's values of teaching practice; planned modifications might address those gaps. However, be sure to also inquire about the timeline for the planned changes.

If you are considering modifying an existing instrument, we recommend that you follow the processes outlined in the previous sections. Modifications, no matter how trivial they may appear to be, can hold consequences for the instrument's usability, reliability, and validity.

Observer Training and Certification

THE FOUNDATION OF CONSISTENT AND ACCURATE OBSERVATION SCORES

Once you have developed or selected an observation instrument, instituting observer training is the next critical step in establishing a valid district-level teacher observation system. The primary goals of observer training are to guide observers' understanding of the dimensions of the instrument and its rubrics and to give them an opportunity to hone their skill in applying the rubrics accurately. Without this step, the promise implicit in a shared definition of teacher effectiveness cannot be realized. To provide consistent and accurate observation scores, all observers must have the same understanding of what constitutes each level of teacher quality the system describes.

MET Project Example

Master Coding

Each of the instruments examined as part of the MET project used a variation of a master-coding process. Instrument developers ran their own master-coding sessions. Four of the instruments—The Framework for Teaching (FFT), the Classroom Assessment Scoring System (CLASS), the Protocol for Language Arts Teaching Observations (PLATO), and Quality Science Teaching (QST)—divided lessons into 15-minute segments, while Mathematical Quality of Instruction (MQI) used 7.5-minute segments. These time segments were chosen for master coding based on recommendations from the respective instrument developers. They were determined to be adequate for collecting sufficient evidence for each dimension. For research purposes, master coders timestamped every piece of evidence that was critical to deriving the score for each dimension.

After master coders individually scored the entire lesson, they used a reconciliation process to agree on the evidence and scores. The number of participants in reconciliation meetings ranged from two to six. When there were only two, the reconciled scores were checked by an independent master coder. Multiple checks and balances along the way helped determine "true scores" for the videos.

ETS content leaders participated in the sessions and ran quality control checks to ensure that evidence, corresponding timestamps, scores, and rationales were accurately recorded. To provide a sufficient number of exemplars for use in rater training and scoring support, experts master coded at least 50 videos for each observational instrument. Master coding consisted of selecting videos that represented a wide variety of teacher performance across all score points in all dimensions and timestamping and documenting evidence of behavior pertaining to each.

Your district's options for training programs are related to your choice of instrument. If you chose a publicly available observation tool, you can use an observer training program that was designed for the instrument you selected or develop your own training. If your district chose to develop its own instrument, designing a locally developed training program is the only option. This section outlines the important components of effective observer training and certification programs.

The process of analyzing and scoring the exemplars is called *master coding*, and in the case of classroom observations, the exemplars are usually captured on videotape. These master-coded videos are the crux of training, certification, and ongoing monitoring.

MASTER-CODED EXEMPLARS

The cornerstone of effective observer training is the use of exemplars scored by experts in the instrument. The process of analyzing and scoring the exemplars is called *master coding,* and in the case of classroom observations, the exemplars are usually captured on videotape. These master-coded videos are the crux of training, certification, and ongoing monitoring. While districts may take a variety of approaches to master coding, we describe the core elements involved in the process.

The master-coding process should produce at least one clear benchmark exemplar for each score level of each teaching dimension covered by the instrument. Depending on the number of scoring levels on each rubric, range-finder exemplars—examples of performance at the high and low boundaries of a score level—can prove helpful in illuminating the variety of performances likely to be seen at a given score level. (Range finders may not be needed when an instrument elaborates five or more score levels, as these rubrics parse distinctions between levels more finely.) The full set of exemplars should span the range of the classroom types that trainees will be observing, and they should also represent the range of grade levels, teacher experience, subject matter, and teacher and student diversity seen in the district. As a guideline, a pool of 50–75 master-coded videos is generally large enough for exemplar and practice selection. This number assumes the entire range of scores and the aforementioned classroom types are represented by the pool of master-coded videos.

The Master-Coding Process

Because one of the objectives of master coding is to determine the "true" or correct score for performance on each of an instrument's dimensions, more than one master coder should analyze and score each video lesson. Using multiple coders increases the chances that the final score will be closer to the "true" score than a score provided by a single coder. One model is to group master coders into pairs to begin the process. In this partner model, two master coders are assigned to each video, but the same two coders do not score all video lessons. The partner model includes a reconciliation session—the process by which two or more coders come to agreement on scores and rationales for scores for particular videos and clips. The expectation is that later in the process there will be several other master coders involved in the selection of exemplar videos used in training and assessing observers, and they will validate the scores.

Breaking up and rotating these partners or teams—even those who work well together—after a predetermined number of videos can help minimize the potential impact of pairs/teams who inadvertently contribute errors in scoring (e.g., they may drift off the scale in unison or share a bias). Without this kind of built-in failsafe, such problems could otherwise go undetected because of the coders' high level of agreement. Rotation also minimizes the impact of groupings in which a partner is more assertive than others, which can create false agreement.

Ideally, the master-coding process proceeds as follows:

- Working individually, each master coder watches a videotaped classroom episode and records the pertinent evidence as guided by the instrument (it is helpful to timestamp this evidence so that it can be located easily during later discussion). Evidence includes the behaviors, quotes, assignments, and other facts that are observed. Observation instruments vary in what they focus on, so the evidence collected would depend on the instrument you are using.
- Each master coder then reviews the documented evidence to confirm that each pertinent piece of evidence is recorded and is associated with the correct dimension or scale.
- Each master coder then assigns a score to the evidence according to the rubric for each dimension.
- After all master coders have scored the lesson, they meet for a reconciliation session to establish the best composite evidence and to determine a consensus score.

We recommend that a videotaped classroom lesson be divided into segments for master coding (e.g., 15-minute segments). Shorter segments make it easier to locate evidence during reconciliation sessions and when identifying example clips to use in training. In training, you want to use shorter segments to balance enough observation time without making training time too long.

Additional Considerations for the Master-Coding Process

Here are some additional aspects of the master-coding process you should consider:

- **Selecting master coders:** Who will be your master coders? There is no ideal profile of a master coder, but there are some traits that are important to consider. How knowledgeable are they about the instrument? Have they had experience observing and rating classroom interactions and teaching practice? Will they be able to score in an unbiased way? Are they able to work collaboratively?
- **Training master coders:** Master coders must be trained in the instrument and in the master-coding process. The goals of the training are to ensure that master coders understand the scales and rubrics, know how to sort evidence into dimensions, can identify pertinent evidence, and can score the evidence accurately. To accomplish this task, a pool of already scored videos is needed, or the trainer must know the instrument thoroughly enough to guide the master coders and debate scores on the fly. The training activities should produce evidence that justifies the trainer's assessment that the master coders are proficient enough in the instrument to begin master coding.
- **Ensuring the success of master-coder training:** What will you do if a trainee does not understand the rubric or master-coding process? Will you retrain them or dismiss them from the task?
- **Monitoring master-coder quality:** Monitoring the quality of the work of master coders begins with their training and continues throughout the entire master-coding process. How will you make sure that master coders remain accurate in every aspect of master coding—identifying evidence, recording evidence, timestamping, scoring, and writing rationales? Who will monitor the master coders? What process will be set up to monitor the master coders? What will you do if a master coder starts to drift or produce less than acceptable quality work? Will you hold retraining? All of these questions need to be considered before you start the master-coding process.

The reconciliation session. During the reconciliation session, master coders assigned to the same video segment share their evidence and scores with each other. First, they discuss the evidence to confirm that it is part of the perceived dimension and that all coders interpreted it the same way. When master coders do not agree about a piece of evidence, they can view a section of the video as a group to reconcile their points of view. (It is here that the timestamps collected during master coding become useful.)

After the coders review all of the evidence, they compare their scores. Matching scores are discussed to confirm that they were given for the same reasons. Nonmatching scores are debated until the coders reach consensus. Finally, the coders write a best-composite rationale for the agreed-on score.

The process continues for all teaching-quality dimensions. If the master coders cannot agree about the evidence and scores for a particular lesson segment, then the video should be given to additional master coders. This second group should follow the above process to individually score the lesson, then be brought into a follow-up reconciliation session with the original coders.

OBSERVER TRAINING

Training observers to use an observation instrument can be a challenging task, especially when most of the learners come to the training with prior experience observing classroom teaching practice. What do you think happens when you show a roomful of experienced educators and administrators the same video of a classroom lesson and ask them to rate it? When the observers have differential professional knowledge, experience, and preferences that influence their focus, interpretations, and judgments, they produce different ratings and value different aspects of teaching. For these reasons, observer training—a goal of which is to get that roomful of educators to agree on a rating—can be as much about encouraging trainees to forget established models of practice and unlearn old habits as it is about learning new definitions and rubrics.

[O]bserver training—a goal of which is to get that roomful of educators to agree on a rating—can be as much about encouraging trainees to forget established models of practice and unlearn old habits as it is about learning new definitions and rubrics.

Effective observer training is critical to establishing the validity of your district's observation system because the *validity* of observation scores cannot be higher than their *reliability* (the extent to which scores generated through the system's use are consistent and accurate). In other words, a district cannot make valid inferences about the effectiveness of its teachers based on the instrument scores unless multiple observers would arrive at the same scores independently and consistently.

Because it is impossible to establish this kind of consistent, reproducible, and accurate scoring without effective training (Johnson, Penny, & Gordon, 2009), this section outlines important considerations for planning observer training for your district. In making the many choices before you, be sure to consult two vital resources when you design your training:

- The instrument developer—to ensure the content is accurate and sufficient.
- Instructional design specialists—to ensure that the training meets the needs of adult learners.

The Training Mode

The mode of delivery you choose for your observer training depends on your setting and needs. Among the factors that could influence your decision are the goals of the training, the number of participants, the quality of the district's trainers, the timeline available for the training (in terms of both duration and schedule), and the participants' desired method of learning. You have essentially three choices as to how to deliver training materials to your observers:

- Face-to-face training
- Online training
- A hybrid of face-to-face and online training

Face-to-face training, a historically popular professional development model, appeals to many instructional designers, trainers, and trainees. In addition to providing opportunities for observers to learn from each other, this mode allows trainers to use formative assessment methods that can result in on-the-spot retraining. However, the use of formative assessment requires strong trainers who are proficient in the instrument and the training goals and can adjust the training on-the-fly. Also, monitoring trainee learning in this manner, while desirable, loses its effectiveness with groups of more than 40 trainees. If you have to train a large number of observers, holding multiple training sessions or using many trainers are two ways to encourage this kind of instructional monitoring. Two disadvantages to face-to-face training are a lack of alternative options for trainees who cannot make all meetings and the many logistical details to be addressed.

Online training, which has become increasingly popular in recent years, offers several potential advantages over face-to-face training. Choosing this mode:

- Minimizes logistical arrangements.
- Standardizes information being conveyed to all trainees (i.e., reduces trainer effects).
- Allows trainees to set flexible training schedules and learn at their own pace.
- Offers trainees a resource for revisiting concepts and video examples.

MET Project Example

Training Modes

For the MET project, some raters were trained online while others received face-to-face training. The majority of the raters were trained online through a series of modules that included bias training and in-depth coverage of each component of the observation protocol they would be scoring. Judging by levels of rater reliability, this training was deemed effective and even presented the additional benefit of the raters being able to review the modules whenever they chose during the scoring process.

A small group of raters received face-to-face training on the Classroom Assessment Scoring System (CLASS) instrument in order for these raters to be able to take part in a pilot study prior to the main MET project. These raters enjoyed the benefit of having direct personal interaction with the trainer, allowing them to ask questions to seek clarification throughout the training process. To account for this during online training, raters were provided with contact information for experts on their observation protocol, whom they could email with questions throughout the training process.

To achieve one of the major pluses of face-to-face training—providing on-the-spot remediation or further explanation when trainees struggle with or misunderstand content—designers of online instruction should consider inserting short, frequent assessment opportunities in the training that provide trainees with immediate feedback.

The hybrid model offers districts an opportunity to weave face-to-face and online training in a way that meets the learners' needs and addresses concerns about quality control and flexibility. Under this model, the majority of the content—particularly content that requires consistent presentation—is placed online, but the trainees come together at the beginning, middle, and end of training for the following purposes:

- **To launch the observation program and training:** This initial face-to-face session allows observers to ask questions and experience some of the training as a group.
- **To assess trainees' growing understanding:** This midpoint get-together provides trainers an opportunity to clarify any confusion about the instrument.
- **To review trainee understanding prior to taking a proficiency assessment:** This final session can provide a means of strengthening assessment outcomes.

With advancements in technology, cameras, and online meeting venues, some of these face-to-face meetings could be accomplished in small groups online. This kind of setting allows trainers to have real-time discussions with trainees based on their individual needs.

Training Activities and Content

This section covers instrument training activities and content. It is important to note that feedback training is not included in this discussion even though it is an important part of developing an observer's skill set. Ideally, feedback training occurs once an observer has demonstrated proficiency in applying the instrument accurately and consistently.

Throughout instrument training, observers need multiple opportunities to practice their new skills. Activities should provide observers with feedback about their understanding so they can make adjustments and hone their abilities. The majority of this practice will make use of your master-coded videos, but activities that assess whether observers can distinguish evidence from opinion statements or can sort evidence into appropriate dimensions can also prove useful.

Frequently Asked by Practitioners

Prioritizing Implementation

Q: What is the best way to deploy training in a nonresearch place such as a school district where you must implement with a number of principals in a limited time? What is the most important part to implement now?

A: The most important thing is making sure the principals have a true understanding of the instrument. Observers must sort evidence into the right dimensions. If they do not do this, they may favor some dimensions and downplay others. Classroom interactions are complex, and simplifying the use of the instrument elements is challenging. The process is about helping people unlearn prior conceptions of teaching practice, as well as learn new ones. However, to get started, you could put lots of power into the dimensions and score points that are more common in your district and fill in the gaps a little later.

[R]ater bias is inherent in all performance assessment; consequently, bias awareness training is an important component of rater training. Training to recognize bias becomes even more important when scoring video and in-class observations because of the wide variations in what an observer may see or hear and thus attune to.

The content objectives for the observers you plan to train are:

- To learn the instrument and understand how it defines quality teaching.
- To learn observation skills.
- To learn how to apply the rubrics and score all dimensions accurately.
- To learn to minimize the impact of professional biases.

The subsections that follow address each of these elements in turn.

The observation instrument. The first step to increasing the reliability of your district's observations is to make sure that all observers understand the instrument. The instrument articulates the aspects of teaching and learning that are valued in your observation system, as well as how these aspects are measured. To instill confidence in trainees, observer training should provide information about the development, validation, and research base for the instrument. Trainees must grasp the structure of the instrument, understand the definitions of dimensions (sometimes also called *components* or *indicators*), and know what evidence to collect for each dimension. Making connections and distinctions among the dimensions can help observers understand how to sort evidence.

Observation skills. All trainees—even those who have been observing classrooms for years—must learn and follow the observation process set out for this instrument if your district is to achieve standardized observations. The trainees must also internalize specific observer skills, such as becoming attuned to words and behaviors, recognizing evidence as a set of facts without opinion, distinguishing key evidence from other evidence, accurately sorting evidence into dimensions, and accurately documenting evidence. Observer training should include activities designed to sharpen these observer skills so your observers learn to use the lens of the instrument to search for and record evidence consistently across classrooms.

Performance rubrics. One of the more challenging tasks of observer training is learning to apply the rubrics. The language of the rubrics may seem clear, but it may blur for the observer when evaluating teaching and learning practice. Practicing with video exemplars can help observers make the transition from reading a rubric to applying it. We advocate the use of video exemplars to promote standardization in observers' training experience. Using videos also provides flexibility in selecting a diversity of lesson samples to which learners may practice applying the rubric. The videos selected for practice should represent concrete examples of the scores for each dimension. Each video segment is accompanied by a list of evidence and a rationale that together explain why it is an example of the score level. Depending on how many score levels are used in the instrument, range finders can help trainees get a clear understanding of the range of performances that may be seen within a score level.

Bias awareness training. Despite rigorous training in the observation instrument, everyone has professional preferences or biases that can influence their judgments about teaching. In fact, rater bias is inherent in all performance assessment; consequently, bias awareness training is an important component of rater training. Training to recognize bias becomes even more important when scoring video and in-class observations because of the wide variations in what an observer may see or hear and thus attune to. Bias

awareness training helps ensure fair and valid scoring by helping observers become aware of their individual biases and how they might inadvertently sway their evaluation of a teacher's performance. Once observers are aware of their own biases, they can be trained to minimize the impact of those biases.

Bias, as the term is used here, refers to factors unrelated to teaching practice (i.e., construct-irrelevant factors) that can influence an observer's scoring decisions. Bias occurs whenever variation in an observer's application of the scoring criteria can be tied to particular characteristics of what he or she is scoring—in this case, for example, a video's setting or the individuals captured in the video—rather than to the quality of teaching as defined by the criteria.

Bias awareness training should address the following objectives:

- To increase observers' understanding of the kinds of biases and professional preferences that can influence the quality of observations of teaching practice.
- To provide observers with opportunities to identify and explore their biases and professional preferences.
- To provide observers with strategies for monitoring and reducing the influence of biases and professional preferences on observations.
- To help observers identify a list of "triggers" for their underlying biases that will assist in maintaining awareness of their biases.

Observers need to clearly understand that biases can affect their scoring. They also need to understand that everyone has biases; that biases can be positive or negative; and that biases are a natural by-product of personal experience, environment, and/or social and cultural conditioning. While many observers readily recognize specific prejudices and stereotypes as biases, we must also help them learn to identify personal preferences and internalized societal and cultural views that can affect scoring. The goal of the training is not to completely erase bias but to at least minimize its effects.

Because, in general, people are hesitant to discuss their personal views, self-reflection exercises can help observers distinguish between professional judgment and professional preference, and they may assist observers in identifying their own bias blind spots. Word association exercises, such as those used in implicit attitude tests and reactions to short video clips, can further assist observers in identifying and reflecting on

MET Project Example

Bias Training

The Bias Training Module designed for the MET project was a required component for all participating observers and was the same for all instruments. While it included checkpoint questions, it did not include an assessment as to whether raters "passed" or "failed" the training.

All training was conducted online and at the rater's own pace. After being given an overview of bias and its impact on video and in-class observation training, observers were led through a series of exercises to encourage them to think about and identify their individual bias triggers. They were encouraged to develop and keep a "trigger list" of features they recognized might cause them to score a video higher or lower due to a personal bias. Exercises included recording word associations and immediate reactions to written and video scenarios.

their biases. While identifying biases is critical, it is also important for observers to keep a list or other record of their bias triggers or blind spots so that they remember what to watch for; recognizing triggers is key to preventing them from influencing scoring.

Here are some questions to consider when developing bias training:

- **Mode of training:** Should it be face-to-face, online, self-directed? While face-to-face training does not always elicit open and honest feedback from participants, some feel online or individual, self-directed training is isolating and does not allow for adequate feedback. As discussed earlier, combined modes may be more effective.
- **Specificity:** Should we target specific biases or treat bias as a more general phenomenon? Districts can develop training that does either or both—elicits reflection on individual biases and/or on specific biases, such as those stemming from variations in regions, districts, environments, socioeconomic conditions, teaching styles, and teacher appearance.
- **Teaching content:** Should the bias awareness training be content neutral or should it be tailored to specific teaching content? Some disciplines or grade levels may elicit a broader or different set of bias triggers. Presence or absence of particular types of lab equipment or the actual lab set-up, for example, could trigger biases for someone in the sciences.
- **Amount of training:** How long should the training be? How many exercises and/or videos should be included in the training? Factors that may influence this decision include attention given to potential biasing factors in other parts of the training, whether training modes are combined, and the extent to which there are known problems. It is important, however, not to give bias awareness training short shrift on the premise that everyone knows about this. The point of the exercises is to make observers aware of biases they may not know they have.

PILOTING YOUR INSTRUMENT

Now that you have developed your training, it is time to evaluate how well the instrument works before full implementation. Whether your district opts to develop an observation instrument from the ground up or to select an available instrument, you should pilot the instrument in your district before using it to make decisions about teacher quality. Do not skip this step, even if research on the instrument supports the reliability and validity of inferences based on observation data collected in other settings. As we stated, validity is a

Frequently Asked by Practitioners

Where to Start?

Q: We have five days total to train principals on our instrument. Where should training start? What speed should we go at?

A: Concentrate on the rubric. Use it to talk about the big picture so people can see where evidence will be collected throughout the whole system. Then, go into the actual scoring of the observations. Provide evidence and have them sort evidence into components, possibly without videos yet. Then, go through benchmark videos, indicator by indicator (for the MET project we tried to cover a few indicators at a time and found that's too large of a chunk).

For five-point rubrics, you could initially show benchmarks for 1, 3, and 5 for each component to anchor raters, and then fill in their understanding with levels 2 and 4. Cover the indicators together afterward. Then, put it all together to connect things.

contextual property. It not only depends on how scores are to be used, but it also depends on the conditions under which scores were collected. And conditions can differ from setting to setting. We recommend the use of videos to conduct your pilot. As stated before, videos promote standardization, and they also give you the flexibility of selecting a diversity of lesson samples to use in piloting your instrument. A pilot can be conducted in a live setting. However, videos will also allow you to evaluate the usability and reliability associated with your instrument across multiple observers more easily and unobtrusively than when the instrument is piloted in a live setting.

Conduct your pilot with observers who were not involved in the instrument development process—those with only basic qualifications and training on your instrument—and with the following aims:

- Determine if observers can use the instrument independently with adequate accuracy and reliability. If these observers have difficulty using the instrument, carefully examine whether the criteria at each score level are clear, objective, and distinct from the criteria at other score levels.
- Be cognizant that relatively high-inference dimensions of teaching practice (e.g., student cognitive engagement) may result in less reliable scoring. You may need to develop specific observer training or augment standard training that is part of an existing instrument to address the challenges of scoring high-inference dimensions. For example, during training you might instruct observers to compare their evidence to an expert's evidence in the high-inference domains. This exercise may show observers exactly what to look for when it comes to collecting evidence for certain dimensions. In addition, you may need to provide observers with more scoring practice and feedback for these dimensions.
- Consider interviewing observers after the pilot to determine whether other aspects of the observation process (e.g., scoring multiple dimensions at the same time) interfered with the quality of their judgments. Their feedback will be invaluable as you further develop observer training.

OBSERVER CERTIFICATION

Observer certification provides the next level of quality control for the observation data you collect with your teacher observation system. It screens observers who have not mastered the ability to apply the scoring rubrics accurately and consistently from those who have done so. Observer certification allows districts to make the following claim: "Observers using the [instrument name] have successfully demonstrated the requisite competency to provide accurate and consistent scores." This section discusses key principles involved in the development of an observer certification test to support that claim.

Observer Certification Test

If you have adopted a publicly available observation instrument, it is likely that an observer certification test for that instrument is also available. However, if using the publicly available observer certification test is cost prohibitive or if it does not measure the full range of observer competencies you deem are important, you must consider developing your own certification test. If you are developing your own observation instrument, you must develop a certification test to assess how well observers are calibrated to the instrument. The following is a high-level description of the aspects of certification test development that you should consider.

The first step in developing a certification test, as with any mastery test, is to determine the competencies that are essential to the job the observer performs. This helps guide the decision about what the observer certification test will measure. However, the certification test should only measure what observers had the opportunity to learn in training.

Generally, certification tests comprise the same kinds of performances an observer encounters in practice. Therefore, observer certification tests often make use of master-coded videos. Ideally, at least one video segment should be included on the test for each score level and dimension represented on the instrument. While it might be nice to include multiple benchmarks for each score level and dimension, this may not be feasible, given the nature of the assessment. It is not uncommon to see certification tests with one video, but we caution against such measures of mastery; a one "item" test will not reliably measure proficiency. (One long video could be used if observers were required to provide scores for each dimension at multiple points in the lesson, as described under the MET project example below.)

The main objective of the certification test is to assess observers' mastery in applying the score scale to a representative set of lessons. In addition to reflecting the full range of performance levels, "representativeness" may be informed by such factors as classroom type, grade level, teacher experience, subject matter, and teacher and student diversity.

Video length depends primarily on the amount of time that is necessary to observe behaviors (a preponderance of evidence) in a given dimension. In addition to determining the length of the certification video, you will need to consider where in the lesson the clip begins and ends. You may use content-expert judgment to determine these points in time, or you may take an empirical approach such as the one used for the MET project. In the latter case, master-coded timestamps and evidence were used to identify the time spans in

MET Project Example

Observer Certification

Each of the observation protocols used in the MET project had a unique certification test, illustrating the point that a certification test should be carefully molded to fit the particular observation protocol. The certification tests differed greatly from one another in terms of the number of videos that comprised them, the length of each video, and the number of components observers scored for each video.

Each of the certification tests was designed to be completed in approximately four hours. For some instruments, such as the Classroom Assessment Scoring System (CLASS), observers were required to take one certification test on the whole instrument. For other instruments, such as Framework for Teaching (FFT) and Mathematical Quality of Instruction (MQI), observers were required to take a separate certification test for each set of three dimensions from that instrument.

Within each certification test, the number of videos that each observer had to score ranged from one to 16. When observers were only required to score one longer video, such as for the Quality Science Teaching (QST) certification tests, they had to provide separate scores for each of the many dimensions depicted at different time points on the video. On the opposite end of the spectrum, the MQI certification test included 16 videos, but each was a two- or three-minute segment from a class. This choice was made because it was pivotal for observers to acknowledge particular activities for this instrument.

Regardless of the format of the certification test, it was important that at least 10 scores were collected to ensure a valid conclusion about the observer's scoring. With all things being equal, the more data points per observer, the more reliable the inferences about his or her performance.

which a preponderance of evidence was recorded for each of the dimensions on a given instrument. For some instruments, the findings of the study were conclusive enough to reduce the amount of time raters were required to watch the video for a particular set of dimensions.

Observers should be given at least two opportunities to pass the certification test, and they should be provided with feedback and an opportunity to remediate their performance between attempts. More attempts, however, require an adequate number of master-coded videos to build parallel versions of the test. Parallel sets are structured so that the representative content (i.e., grade level, content areas) and the level of difficulty are equivalent.

Difficulty is a statistical attribute of the videos and can only be ascertained once you have collected some data from observers. However, you can control some factors related to video difficulty during master coding:

- **Video quality:** Videos that distort or do not present students' facial expressions with high fidelity make it hard to score dimensions that require observers to make inferences based on nonverbal behaviors. Such videos should be excluded from certification sets.
- **Audio quality:** Observations depend equally on what teachers and students say and do. Videos that distort or diminish voice quality should also be excluded from certification sets.
- **Borderline cases:** While borderline cases—videos in which performance falls on the border between two score points—are excellent for training, they are not suitable for certification.
- **Nuanced cases:** Nuanced cases—videos that have unusual characteristics that may be unfamiliar to or problematic for observers—should be flagged for training, not certification.

Score Rationales

In addition to scores, you may decide to collect other data to measure the extent of an observer's skill. For example, you might also choose to collect observers' score rationales for a subset of the observations included on the certification test to ensure that the observers provided accurate scores for the right reasons and not by

MET Project Example

Score Rationales

The certification tests used for the MET project only measured whether observers were able to supply accurate scores for the videos and did not determine whether the scores were provided for the correct reasons. The main purpose of the MET project video scoring was to collect reliable and accurate data, which differs slightly from the purpose of typical observations in practice.

However, to guarantee that all MET project observers had highly qualified scoring support throughout the project, we provided a stricter certification test to scoring leaders—individuals who oversaw the scoring quality of the observers. This certification test asked scoring leader candidates follow-up questions about the scores they provided. The questions asked candidates to either justify why they chose a score or explain why the score was more appropriate than an adjacent score.

Although requiring justifications may have made the tests more difficult to score, in this case, we felt it was important to elicit evidence that the scoring leaders actually understood the thinking behind using the instrument properly in order for them to provide mentoring to MET project raters who may have had difficulty understanding certain score points within particular dimensions.

guessing or by chance. This evidence demonstrates that the observer has internalized the values of teaching implicit in the instrument. Since most observers are responsible for conducting feedback discussions with teachers, evaluating their production of scoring rationales is very reasonable.

Here are some things to look for when examining score rationales:

- A clear understanding of the scoring rubric, accurate appropriation of evidence, accurate sorting of evidence, and accurate understanding of the assigned score level.
- Signs of bias or personal preference, such as comments to the effect of "I would have done it this way …" or "I would have preferred if the teacher told the students … ."
- Language, evidence, or ideas that signal the observer is relying on a competing framework.

Keep in mind that the evaluation of score rationales requires having enough staff members who are sufficiently trained to evaluate and handle the volume of score rationales that the district generates.

Setting Certification Cut Scores

The objective of judgment-based standard setting is to determine a cut score, or passing score, that is reasonable—that is, a cut score that content or subject-matter experts would agree is reasonable. The cut score represents the observer who possesses the minimum competencies and skills necessary to perform as an observer in the field. However, in lieu of data-driven standard setting, an initial threshold for minimal observer proficiency can be established through content-expert judgment. Bear in mind, however, that the purpose of standard setting is not to guarantee that future test-takers pass.

Picture a target with a bull's-eye in the center. Imagine that the rings on this target represent all possible levels of performance for the dimension of teaching practice being assessed through observation and that the bull's-eye represents the "true" level of a teacher's performance that the observer must hit. While we might not expect the observer to hit the bull's-eye with consistent precision for each observation score, we can set an expectation for *acceptable* accuracy and consistency that is commensurate with the conditions of the observation design (e.g., the length of the score scale, the extensiveness of training, and the clarity of the scoring rubric).

MET Project Example

Certification Cut Scores

The Classroom Assessment Scoring System (CLASS) is a content-neutral measure of teaching practice that assesses 12 dimensions of teaching practice. These domains are scored along a seven-point score scale. CLASS developers established an initial certification cut-score of 80 percent exact plus adjacent agreement across scores provided for a 15-minute video clip. The cut score was later revised to 70 percent exact plus adjacent agreement to increase the pass rate (see "Balancing multiple considerations," p. 22).

The Framework for Teaching (FFT) is also a content-neutral measure of teaching practice. For eight Components in Domains 2 and 3, which are assessed along a four-point score scale, the developers established a cut score that required at least 50 percent exact agreement and no more than 25 percent discrepant scores (defined as being two or three score points away from the "correct" score) across scores for two 15-minute video clips. Considering the length of the score scale, this was a fairly stringent standard to meet.

Prior to deciding on a cut score, we strongly recommend that a panel of content experts define the minimally proficient observer while considering the following:

- What overall level of **"exact agreement"** is the minimally proficient observer expected to exhibit in practice? In other words, how often do you expect the score this observer assigns for a performance to *match* the "true" score assigned by a more experienced observer?
- What overall level of **"adjacent agreement"** is the minimally proficient observer expected to exhibit in practice? That is, how often do you expect the score this observer assigns for a performance to be *near* the "true" score assigned by a more experienced observer (depending on the number of levels on the scale, "near" may be defined as one up or down)?
- What overall level of **"discrepancy"** is the minimally proficient observer permitted to exhibit in practice? Or, how often do you expect the score this observer assigns for a performance to be off by more than an adjacent level from the "true" score assigned by a more experienced observer?
- What **overall** level of exact, adjacent, and discrepant scores is the minimally proficient observer expected to have for each of the dimensions of performance being assessed? In other words, are there certain dimensions of performance for which the minimally proficient observer is expected to have lower accuracy than others (e.g., high-inference dimensions)?

Bear in mind, however, that the purpose of standard setting is not to guarantee that future test-takers pass.

Agreement statistics. Agreement statistics capture the extent to which an observer's scores match the master-coded or "correct" scores for the dimensions of teaching practice assessed by the certification examination. "Percent agreement" is commonly used to measure overall observer performance. "Percent exact agreement" tells us how often an observer applies the scoring criteria accurately—in other words, how frequently he or she hits the bull's-eye—across the dimensions of the instrument.

"Percent exact agreement plus adjacent," on the other hand, combines the degree of accuracy and near accuracy the observer demonstrates on the certification test. Exact agreement plus adjacent may be a reasonable gauge to use if there is sufficient conceptual justification for the decision; it is usually applied when the score scale is longer than four points, conceptual boundaries between score points are close, the response or stimulus is relatively complex, and support mechanisms are in place to help observers learn to discriminate between performance levels more consistently and accurately.

We recommend that a minimum percentage of exact scores be required if you use an exact-plus-adjacent performance measure. An observer who passes certification with all adjacent scores has a much different level of proficiency than an observer who passes with all or most exact scores. A high percentage of adjacent scores can be an indication that the observer has not fully learned how to discriminate between levels of performance.

A high percentage of adjacent scores could also signal "playing it safe" behavior. For some observation instruments, the distribution of performance tends to be normal; that is, scores cluster in the middle score categories. Under testing and monitoring conditions, observers who understand this phenomenon of current teaching practice may choose to award scores in the middle of the distribution (right or wrong), knowing that

by and large those scores will either be exact or adjacent with the master codes, and therefore will not raise any flags and prevent them from conducting live observations. Examining the percentage of exact scores, particularly at the tails of the score distribution, can help detect this type of behavior.

A hybrid measure that combines some of these agreement features is another option. You could determine a minimum standard for exact or exact-plus-adjacent agreement in combination with a maximum number of allowable discrepancies. This measure provides a more accurate picture of an observer's performance.

Standardized differences. Standardized differences between an observer's scores and master-coded "true" scores can also be used to establish a cut score. There are several ways in which one can develop a standardized difference value. This section describes one approach that is especially useful when the scale length varies across the dimensions of the instrument, as was the case with one of the instruments in the MET project, Mathematical Quality of Instruction (MQI).

Standardizing the difference between an observer's scores and the master codes establishes a common scale by which to compare performance across the dimensions in the scoring rubric. It is not uncommon to standardize the average difference scores within a dimension by the width of the score scale. Doing so puts the average difference score across all the dimensions on a common scale that ranges from 0 to 1. The closer the sum of the average standardized difference is to 1, the less accurate the observer. If training were effective, you would expect the standardized difference score to be relatively small.

Balancing multiple considerations. Often in standard setting, multiple factors influence the decision about where to set the cut score. In addition to considering the face validity of the cut score, the choice of cut score may have to be balanced against policy concerns or practical matters. The consequences or impact of a particular cut score may have to be weighed.

A more stringent cut score will allow fewer observers to pass through into live classroom observation. However, it will also provide more confidence that those individuals who do pass through possess the requisite competencies and skills to provide accurate observation scores to teachers. Just as important, a more stringent cut score will reduce the chances that an observer who is not proficient will perform live observations.

On the other hand, a more liberal cut score will admit more observers through the doorway into live classroom observation. However, it may also increase the likelihood that observers who have not completely mastered the skills of observation have joined the ranks. In this case, sufficient support mechanisms and monitoring should be put in place to ensure the accuracy of the observation scores assigned to teaching practice.

MET Project Example

Implications for Changing Cut Scores

For the MET project, the decision of where to set the cut score had to be contemplated in light of the volume of videos that needed to be scored in time for the project to carry out its analyses of those scores. To accomplish this, the certification cut scores were lowered in some instances to allow more raters to pass through to operational scoring.

The impact of this decision was an increase in the number of raters who needed more support during calibration (a post-certification consideration that is discussed in more depth later in the paper). MET project raters had to pass a calibration test and receive coaching from scoring leaders (depending on their calibration test performance) before they could score any of the study videos.

Validity Check on Location of Cut Score

Once the initial cut score has been set, it is important to gather evidence to support the "reasonableness" of its location. One way to do that is to examine the performance of typical observers who complete your training against the scores generated by third-party audits (i.e., audits conducted by independent external observers from outside the district), if those data are collected by the observation system. Your examination should include observers who were above and below the certification cut score. You would expect most of the observers who were further below the cut score to be at the lower end of the third-party score distribution. You would expect most of the observers who were further above the cut score to be at the higher end of the third-party score distribution. If not, this would suggest that the cut score was too conservative and allowed in individuals who did not possess the appropriate level of scoring accuracy.

However, it is important to keep in mind that when using an external criterion, such as observations by third-party auditors, it is essential that the unreliability associated with the criterion be taken into account. External observations must be conducted by observers with demonstrated proficiency and accuracy. If the criterion cannot be trusted, then the inferences made based on the alignment between observers' accuracy and the criterion cannot be trusted.

Consequences for Observer Performance

What happens to observers who do not pass the certification test? For large-scale assessment and research, and for most content areas, fairly large pools of qualified raters exist; raters who do not pass certification are simply not invited to score for the program or study. What happens in a small school district that relies on a few classroom observers to evaluate the entire staff? The pool of available observers is not as inexhaustible. In such cases, the stakes for passing certification can be high.

MET Project Example

Consequences for Noncertification

For the MET project, much consideration was given to decide how to handle observers who did not successfully complete the certification examination on their first attempt. Since a limited observer pool was available and these observers may have not scored far below the target cut score, observers who failed to pass the certification examination were asked to further review their training materials, which included several practice videos. They were then given a second opportunity to take a parallel form of the certification examination. Observers were not informed about dimensions for which they were discrepant on their initial attempts; it was decided that it was preferable for observers to refresh their knowledge on the entire instrument before taking the second examination.

If observers passed the second examination, they were considered eligible to score during the MET project. Although the possibility that an observer only performed better by chance on the second examination was a concern, the MET project used multiple monitoring measures throughout the scoring process to ensure the observers were truly scoring accurately, which alleviated this concern. Because using certified observers helps ensure the data collected on each teacher are reliable and valid, observers who did not successfully complete the certification exam on the second attempt were not allowed to score during the MET project.

We highly recommend that uncertified observers do not provide observation scores for teachers until they have demonstrated proficiency. Here are some approaches you can take to help uncertified observers reach the desired level of proficiency before they take another certification test:

- Assign a scoring expert as a coach to guide the observer through the areas of the instrument and rubrics that he or she does not understand well or applies inaccurately.
- Have the coach explore the possible influence of bias and personal preferences with the observer.
- Assign a certified observer to pair up with the uncertified observer during live observations.

REPORTING CERTIFICATION DATA

After a certification test is given to observers and scored, districts must make important decisions as to how much of the resulting certification data they should release and to whom. Districts have many different reporting options here and a case can be made for and against reporting each piece of data described in this section. Generally speaking, when you decide to release information, always report the results in context. Provide information about the kinds of items that are included on the certification test (e.g., videos, multiple-choice items, or mixed-format questions), the mode and extensiveness of observer training, and instrument characteristics (e.g., number of dimensions and score points).

Although we do not recommend releasing information about individual performance, releasing summary statistics about observer certification scores may provide credibility for the district's observation system or for particular observations that are made.

Pass-Fail Status

While this result must be released to observers so that they know whether they passed the certification test or need to take it again, a more troubling issue can arise when an observer fails. If as a consequence of this status the observer is asked not to perform any observations until their next attempt at the certification test, this outcome, even though not published, will eventually become known to teachers.

Actual Scores

Releasing actual score information to observers will allow them to respond appropriately to their scores. For example, if they barely passed the test, they can review their training materials before proceeding with observations. On the other hand, observers who perform well may be tempted to share or discuss their answers with colleagues who plan to take the test in the future. If the district created or purchased multiple versions of the certification test, this may not be an issue, but given the cost of identifying and master-coding videos for this purpose, many districts will not have an adequately large item pool and this may lead to problems with the integrity of the items used.

MET Project Example

Reporting Certification Results

For the MET project, all certification tests were scored by ETS. Candidates were notified only of whether they passed or failed the examination. They were not informed of their scores, which components they scored accurately, or which components they failed to score accurately. This decision was made because a limited number of certification videos were available for the study and it was important that the answers for the items not be compromised.

Additionally, districts may choose to release this information to the public, especially if a large percentage of observers are doing well on the certification test. Although we do not recommend releasing information about individual performance, releasing summary statistics about observer certification scores may provide credibility for the district's observation system or for particular observations that are made. Sharing this information may also help teachers understand how well trained the district's observers are in the observation instrument.

Performance on Individual Dimensions

Releasing information on how candidates perform on individual components can help observers identify components they were discrepant on, which they can then further review before making classroom observations. Additionally, knowing which components they are accurate on and which they need to gain a better understanding of can help observers who need to retake the certification examination focus their remediation efforts. On the other hand, when observers are told that they need to improve on a single component, they commonly only review that particular component, and on their next attempt at the test, they often do poorly on other components. Similar to the release of actual scores, releasing information about performance on individual components can also raise further questions about raters' sharing information and compromising the test items.

Additionally, since most certification tests only contain between one and three items for each component, data about an observer's performance on an individual component are not very reliable. Thus, districts should warn observers to carefully interpret this information if they decide to distribute it.

Responsible Use of Certification Data

Regardless of which certification data districts distribute to observers or the public, all involved parties should be encouraged to interpret and use that data responsibly. The issue is understanding which inferences can be made validly from these score results. The direct interpretation of passing outcomes is that the observer

Frequently Asked by Practitioners

If Someone Fails, Then What?

Q: "What can we do if a principal fails? If there is only one principal, especially in a rural area, and they fail, what can we do? How [do we] proceed?"

A: We have been talking to other districts and states and have found that this is one of the largest shared concerns. Some districts have developed contingency plans. One is to send another observer in with the principal so two "live" observers are in the classroom at the same time. Another option is to videotape the lesson so the principal observes live, and a second observer can watch the video later. The video also gives struggling observers an opportunity to check their thinking about the performance or to view and discuss the performance with another trained observer. These contingency plans should be used until the principal becomes certified.

Another concern this question raises is whether legislatures require or recommend instrument certification. A lower step may be a certificate for a principal who is making an effort toward training but is not certified. In some cases, principals can make observations before passing, but only observations conducted after certification count. An additional option is to consider ways to provide ongoing support to principals who have not passed, such as by providing coaching and opportunities to complete observations with a colleague.

demonstrated proficiency by meeting the established cut score on a set of items. However, this statement is of little interest to most people; even districts will likely want to make additional inferences about observers from their data.

The list of questions that follows points to various conclusions people may want to make based on your certification data. Although they are often necessary inferences, each district should carefully consider whether scores from the certification test can be used to make the inferences validly. One way to proceed is to determine the conclusions the district wants to draw, then look to the district certification test to determine whether it provides evidence to justify the claims. This process may help districts avoid conclusions that are not warranted based on the data collected. Some possible questions include:

- Is this observer qualified to score teachers from all grades of students?
- Is this observer qualified to score teachers from all subjects?
- Is this observer qualified to make live teacher observations?
- Is this observer qualified to just score teacher performance or to provide feedback as well?
- How long will the observer be qualified to provide scores before recertification is required?

Beyond Training and Certification

Human judgment is prone to error and can shift over time. Observers who demonstrate accurate observation skills at one point in time can become more or less accurate over the course of performing observations. Therefore, it is important to implement safeguards to monitor observer behavior and ensure the quality of the district's teacher observations is maintained. This section discusses observer effects such as familiarity bias and other observer effects that can influence the quality of observation data. It also discusses procedures that districts can institute to strengthen the quality of observation systems.

FAMILIARITY BIAS

Familiarity bias is difficult to address and control in practice. It is hard for observers to be objective when they have a vested interest in the success (or failure) of the teachers they are observing. The personal and professional friendships that observers have with the teachers they are observing can lead to more lenient or more severe scoring. The nature of the previous experiences between these colleagues, whether positive or negative, can affect perceptions in ways that may or may not be obvious to the observer.

The greater the level of objectivity and transparency in the observation process, the more confidence teachers are likely to have in scores and feedback. Therefore, we suggest implementing one or more of the following procedures to guard against familiarity bias:

- Instruct observers to acknowledge, prior to an observation, the extent to which their professional or other close friendship with the teacher may influence their observation (i.e., the collection and interpretation of evidence, their scores, and feedback).
- To the extent possible, have an independent within-district observer available if a principal feels he or she cannot provide an unbiased observation.
- Build in quality checks by capturing a subset of the observations on video and having a second person, or team of observers, review them for accuracy.

Frequently Asked by Practitioners

Familiarity Bias

Q: Our observations do not reflect one snapshot of time, but rather a teacher on the whole, over many 20-minute observations, in addition to their interactions with the community and outside the classroom. Because of the nature of our system, we want one person giving all the scores for a teacher, looking at the totality of evidence.

This does not allow easily for outside perspectives. We are concerned, though, about the familiarity bias, so we encourage peer or teacher leaders to sit with the teachers to provide an extra set of eyes or do inter-rater reliability checks, but the principal makes the final decision about the scores. Do you have any suggestions on how to minimize familiarity bias?

A: In this model, it is important to make sure principals and other observers focus only on evidence and that they monitor the influence of their professional preferences and relationships with the teachers they observe. Then, afterward, it's perfectly fine to synthesize the data across the observers.

OTHER OBSERVER EFFECTS

Monitoring, identifying, and minimizing other observer effects can also strengthen the defensibility of your teacher observation system. Other observer effects that can influence the reliability and validity of score interpretations include:

- **Drift.** Drift is a shift in the overall direction of an individual or group of observers' scoring over time toward greater leniency, severity, or even accuracy. Drift affects score comparability. Individual drift can be detected by periodically assigning observers to score a set of observations that has been vetted by an expert. Group drift can be identified through a procedure called *trend scoring*. Trend scoring is achieved by having observers score a set of observations that were scored by the same or an equivalent group of observers in the past.

- **Halo and fatal-flaw effects.** Halo and fatal-flaw errors are viewed as trait carryover effects in which one salient trait or feature of an observed teacher influences the observer's overall judgment. In instances when the domain comprises related but distinct traits, the observer who is influenced by these effects fails to judge individual traits based on their own merit. The term *halo effect* is used when observers make this error in the direction of higher than deserved scores; the process follows that, because A is good, then B and C must also be good. By contrast, with the *fatal-flaw effect*, lower than deserved scores are awarded; because A is weak, B and C must also be weak.

- **Central tendency effect.** The scores of observers who have a central tendency tend to "hang out" in the middle of the score scale—not because all performances warrant a middle score point but out of the observers' reluctance to award score points at the tail ends of the score scale. Possible causes of central tendency include:

  - An observer's lack of confidence that he or she knows how to recognize performance at the tails.
  - An observer's fear of the consequences of awarding such score points, such as an effect on school culture.
  - An observer's belief that the performances do not exist or are extremely rare.
  - An observer's distorted perception of these performances as "fairytale" performances because they bring to mind such extremes.
  - An observer's decision to "play it safe" because scores in the middle tend not to raise red flags.

SUPPORTING AND MONITORING OBSERVERS

Observation Coaches

One way to monitor the quality of your district's teacher observations is to assign observation coaches to monitor and mentor observers. Prior to identifying coaches, you should delineate the scope of tasks they will perform and the amount of time they are expected to devote to these tasks. The coaches you choose to fill this role should have:

- Demonstrated expert-level observation skills.

- The ability to effectively communicate rationales supporting score decisions to observers.

- An understanding of how to motivate observers to put aside personal scoring standards and adopt the standards ascribed by the scoring rubrics.

The process of identifying observation coaches might begin by selecting high scorers on the certification test (e.g., the top 10 percent) because these individuals have proven that they score accurately. If providing score rationales is not a component of your certification test, then you will have to develop a screening test for this capacity. The test can be a short set of master-coded video clips for which "true" scores and expert rationales exist. The observation coach candidates would score and provide written score rationales for these videos. Their rationales should say why the videos received the assigned scores, as well as why the videos did not receive the adjacent scores.

Some one-on-one time with each of the candidates can help you gain a sense of their willingness to be an observation coach and their understanding of the requirements of the role. Keep in mind that some candidates you identify, while excellent observers, may not possess the skills to effectively communicate with and coach others.

Once selected, observation coaches require training. This training may consist of activities that familiarize the coaches with frequently encountered errors and gaps in understanding demonstrated by observers (such as the observer effects described in the previous section), ways to address issues of bias, and protocols for passing along performance information to administrators. Training should also include opportunities for coaches to see appropriate coaching behavior modeled.

Finally, be prepared to provide ongoing support and development for observation coaches. This might include periodic meetings that provide an opportunity for them to discuss issues encountered with observers or to make recommendations for improving observation procedures. Because observation coaches are usually the frontline in an observation system, they tend to have better insight into the day-to-day challenges involved with observation than do administrators. Ongoing development might also include periodic calibration to ensure their observation skills remain attuned to the instrument's scoring levels.

Performance Data

The effectiveness of the system you put in place to monitor the quality of your teacher observation system is driven by the kind of performance data you collect. Two kinds of performance data can be practically gathered for this purpose: (a) calibration data and (b) double-score data.

Calibration Data

Calibration is an ongoing process designed to ensure that observers continue to score accurately. Calibration data are collected periodically by reassessing observers, usually just before they assign scores in a consequential setting (i.e., where the scores mean something and have impact). Calibration assessments, while similar to the certification test, are shorter and may measure a narrower skill set (e.g., apply the scoring rubrics and assigning an accurate score). Calibration also makes use of master-coded videos. Calibration tests should not be time intensive; they should be sufficient to assess whether observers are "ready" to provide observation scores to a teacher or whether their understanding of the scoring levels has shifted. Observers who do not meet calibration performance standards should be given an opportunity for refresher training and to discuss their performance with an observation coach.

Double-Score Data

Double-score data are used to compare two observers' scores of the same performance and can be obtained in a number of ways. For example, as mentioned earlier, some of an observer's live observations may be videotaped for another observer to score. Alternatively, a subset of an observer's scheduled observations can be randomly selected for live pairing. The paired observer can be a randomly assigned observer from within the school or from another school, or he or she can be an observation coach. The subset of independent scores assigned and evidence collected by the paired observer or observation coach provides the basis for measuring the accuracy of an observer's scoring.

The structure of the paired observations determines the validity of any interpretations made about observer performance. For the best result, consider pairing each observer with different observers over the course of the year. If time and opportunity permit, consider pairing observers from different schools with one another. Observers within a school may share more observation behaviors or biases than observers across schools. You may also find that observers within a school tend to agree with one another at a high rate but perhaps for the wrong reasons.

Concluding Thoughts

Two major assessment measurement principles were invoked in this paper—reliability and validity. Reliability and validity are technical terms. But when educators think about a defensible observation system they think in terms of *fairness,* fairness of the observer, the instrument, and the timing of the observations—which can be addressed by employing the practices described in this paper. Those who are charged with implementing and monitoring the observation system know more than anyone how critical it is to get this right. The focus of this paper is on sharing our knowledge about important processes to incorporate into teacher observation systems in order to assist districts in improving the reliability and validity of observations. This is by no means an easy task, but it is doable with the right attention and resources.

REFERENCES

Baldwin, D., Fowles, M., & Livingston, S. (2005). *Guidelines for constructed-response and other performance assessments.* Princeton, NJ: Educational Testing Service.

Bell, C. A., et al. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*(2–3), 1–26.

Johnson, R. L., Penny, J.A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks.* New York: Guilford.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319–342.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kane, T., & Staiger, D. (2012, January). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Seattle, WA: Bill & Melinda Gates Foundation.

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387–431). Westport, CT: American Council on Education/Praeger.

McClellan, C., Atkinson, M., & Danielson, C. (2012). *Teacher evaluator training & certification: Lessons learned from the Measures of Effective Teaching project.* [White paper]. Teachscape. Retrieved from http://www.teachscape.com/resources/teacher-effectiveness-research/2012/02/teacher-evaluator-training-and-certification.html.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice,* 5–8.

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 Jeff Raikes and Co-chair William H. Gates Sr., under the direction of Bill and Melinda Gates and Warren Buffett.

For more information on the U.S. Program, which works primarily to improve high school and postsecondary education, please visit www.gatesfoundation.org.

BILL & MELINDA GATES *foundation*

www.gatesfoundation.org