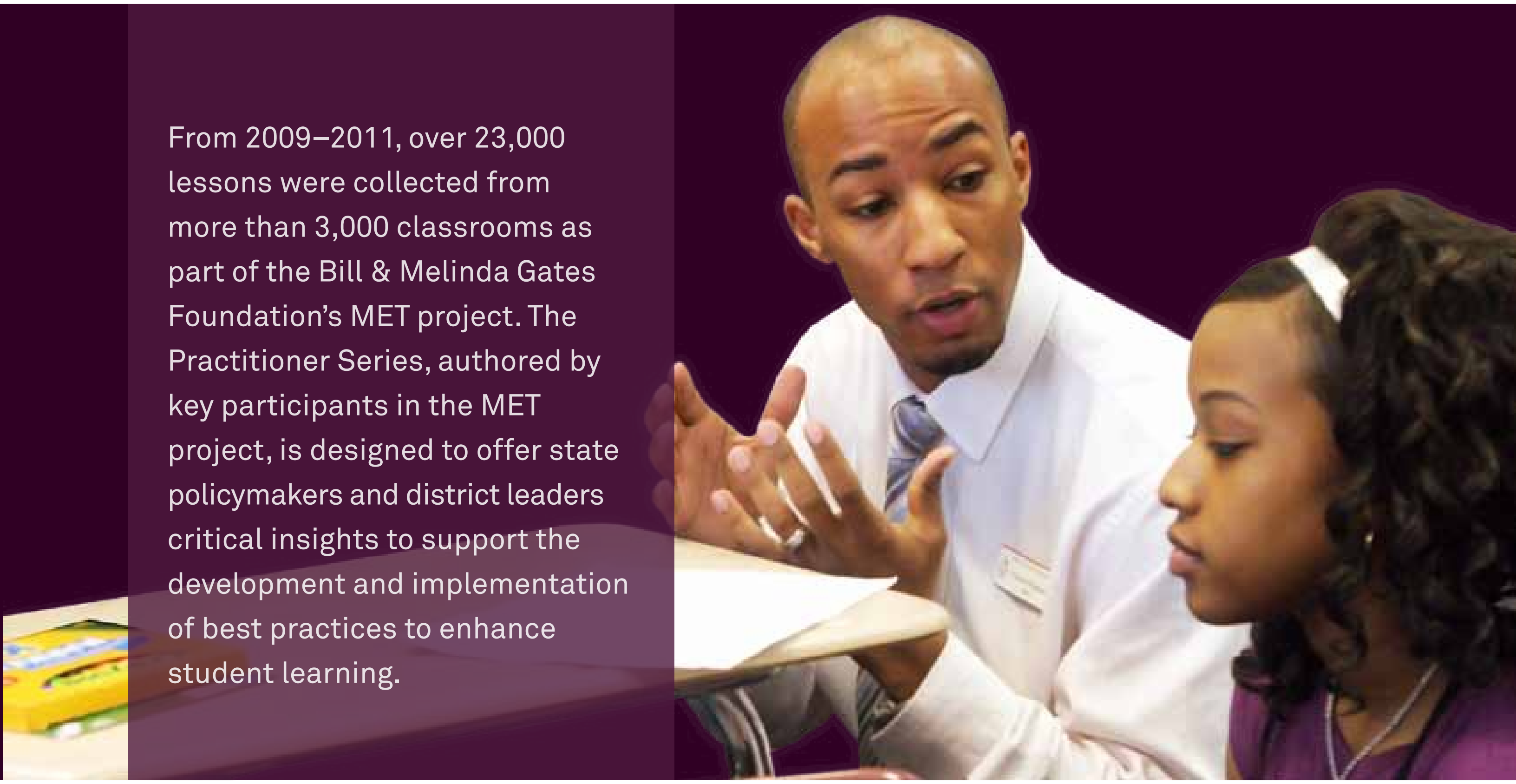
PRACTITIONER SERIES FOR TEACHER EVALUATION

Teacher Evaluator Training & Certification:

Lessons Learned from the Measures of Effective Teaching Project



From 2009–2011, over 23,000 lessons were collected from more than 3,000 classrooms as part of the Bill & Melinda Gates Foundation's MET project. The Practitioner Series, authored by key participants in the MET project, is designed to offer state policymakers and district leaders critical insights to support the development and implementation of best practices to enhance student learning.

AUTHOR

Catherine McClellan | Principal Scientist, Clowder Consulting

CONTRIBUTING AUTHORS

Mark Atkinson | Founder & Chief Strategist, Teachscape

Charlotte Danielson | Educational Consultant and Author

FOREWORD BY

Terry Holliday | Kentucky Commissioner of Education

Foreword

Over the past few years we have seen increased research and action related to the effectiveness of teachers in improving student learning outcomes. States and school districts have invested significant resources in developing teacher effectiveness and evaluation systems. Race to the Top grants and No Child Left Behind waivers have required that teacher effectiveness and evaluation systems be developed and implemented with fidelity.

Having served at every level of educational leadership over the past 40 years, from teacher to state commissioner, I have experienced all sides of the teacher effectiveness and evaluation debate. Teachers want quality feedback that helps them improve their craft, but they're concerned about the fairness of any evaluation system. They're also concerned about the knowledge and ability of the rater who conducts the evaluation and feedback process. Principals and evaluators want to support their teachers, but they are very concerned about the time required for training and conducting evaluations. Superintendents want to support their schools and are concerned about all of these issues, but also about the reliability and validity of the evaluation system due to potential legal issues. Commissioners of education want a quality system in place at all levels and share the concerns at each level, but they are also working to balance all stakeholder groups, including legislators and governors.

This paper addresses these very real concerns faced by educators and policymakers and offers a practical and positive response. Readers will benefit from the excellent insight into these challenges and will find that there is exciting potential for training and for implementing these systems using creative and technology-based solutions.

If we are to improve student learning outcomes in this nation, our teachers need our support and coaching. While a clearly defined effectiveness rubric is an excellent first step, it is only the beginning. The effectiveness rubric will be no better than the evaluators who use the rubric to provide feedback, coaching, and support to classroom teachers. This paper addresses the most critical challenges of training and certification of evaluators of classroom teaching through a well-researched and well-thought-out solution. While much work remains to be done in implementing this type of system, this is an excellent thought process for states and districts to utilize as they develop and implement new teacher effectiveness and evaluation systems.

— Dr. Terry Holliday, Kentucky Commissioner of Education

About the Authors

Catherine McClellan

PRINCIPAL SCIENTIST, CLOWDER CONSULTING

Catherine McClellan is a principal scientist at Clowder Consulting, LLC. Prior to forming Clowder Consulting in September 2011, Catherine spent 13 years at Educational Testing Service, where she was most recently the director of human constructed-response scoring.

While at ETS, Dr. McClellan served as the R&D project director for the Video Plus Scoring project, part of the Measures of Effective Teaching study, funded by the Bill & Melinda Gates Foundation. Dr. McClellan's team took instruments designed for the observation of classroom sessions by experts and modified the scoring design so that they provided data for a statistical model, along with other components to inform inferences on teaching quality and value-added models. Dr. McClellan had responsibility for the design and execution of all preliminary studies, the design of the operational scoring, and the statistical monitoring of the live scoring.

Earlier in her tenure at ETS, Dr. McClellan served as the director of psychometrics and worked on the National Assessment of Educational Progress (NAEP), a federal survey assessment of education in the United States, also known as "The Nation's Report Card."

Dr. McClellan is affiliated with the Psychometric Society, the National Council on Measurement in Education (NCME), and the American Educational Research Association (AERA) Division D. She received her Ph.D. in research and evaluation methodology, her M.Ed. in secondary mathematics education, and her B.Sc. in pure mathematics from the University of Florida.

Mark Atkinson

FOUNDER & CHIEF STRATEGIST, TEACHSCAPE

Mark Atkinson is the founder of Teachscape, Inc. For 11 years, following Teachscape's inception in 1999, Mr. Atkinson served as the company's chief executive officer. He has raised more than $50 million in venture and philanthropic funding for Teachscape and secured groundbreaking partnerships with the Bill & Melinda Gates Foundation, Stanford University, SRI International, the New Schools Venture Fund, Intel, McGraw-Hill, and the American Federation of Teachers. Mr. Atkinson has also served as an advisor to the Bill & Melinda Gates Foundation's Measures of Effective Teaching project. He currently steers Teachscape's long-range corporate strategy with a specific focus on innovative partnerships to expand Teachscape's role in the preparation, licensing, assessment, and development of K–12 teachers. Mr. Atkinson also advises federal, state, and local policymakers on new approaches to teacher evaluation and licensure. He is a frequent speaker at education conferences across the United States.

Mr. Atkinson has served on the boards of trustees of the Oracle Education Foundation and the Breakthrough Collaborative, and is a current director of PresenceLearning, Inc., the country's leading provider of online speech therapy services.

Prior to founding Teachscape, Mr. Atkinson was the senior producer and manager of new markets for CBS News Productions in New York City, and before that he served as a producer for *Peter Jennings Reporting* at ABC News, where he produced a series of Emmy award-winning network specials on U.S. foreign policy in Bosnia, Haiti, and Iraq. He is a recipient of the Alfred I. duPont-Columbia University Gold Baton, considered the industry's most prestigious honor, for his work in Bosnia. Mr. Atkinson is a graduate of Yale University.

Charlotte Danielson

EDUCATIONAL CONSULTANT AND AUTHOR

Charlotte Danielson is an internationally-recognized expert in the area of teacher effectiveness, specializing in the design of teacher evaluation systems that, while ensuring teacher quality, also promote professional learning. She advises state education departments and national ministries, both in the United States and overseas. She is in demand as a keynote speaker at national and international conferences, and as a policy consultant to legislatures and administrative bodies.

Ms. Danielson's many publications range from defining good teaching (*Enhancing Professional Practice: A Framework for Teaching,* 2007) to organizing schools for student success (*Enhancing Student Achievement: A Framework for School Improvement,* 2002) to teacher leadership (*Teacher Leadership That Strengthens Professional Practice,* 2006) to professional conversations (*Talk About Teaching! Leading Professional Conversations,* 2009) to numerous practical instruments and training programs (both onsite and online) to assist practitioners in implementing her ideas.

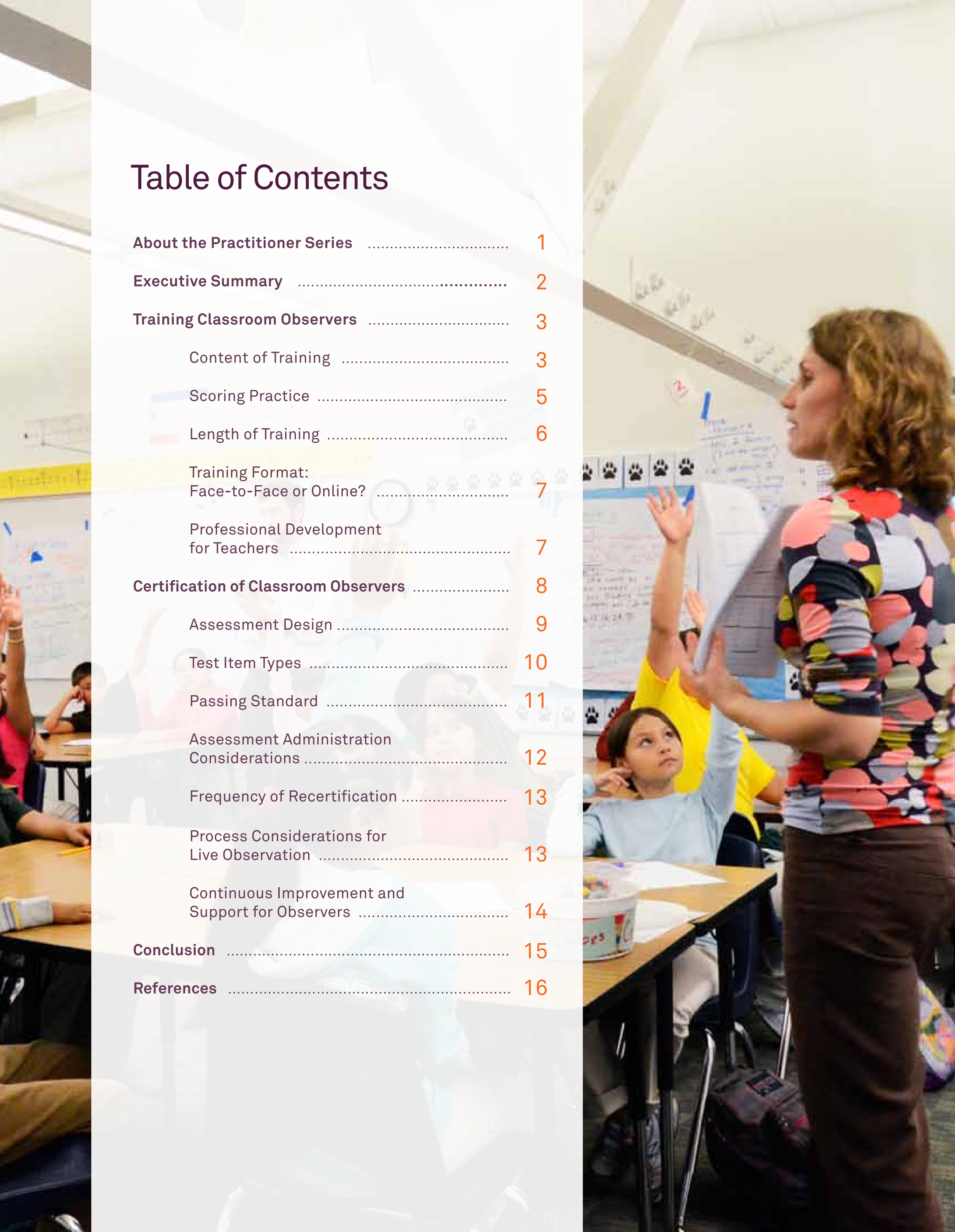


Table of Contents

About the Practitioner Series

State and district leaders know that effective teachers can have a transformative impact on student success. They appreciate that more than any other variable within a school, effective teaching can improve educational outcomes critical to our most pressing education policy objectives: building the STEM pipeline, ensuring that students are reading on grade level by grade three, increasing graduation rates, and ensuring college readiness. For school and district leaders, as well as governors and state legislators, college readiness and completion are not just educational aspirations; they are economic development imperatives. The process of defining, measuring, and promoting teacher effectiveness has become a critical policy lever to achieve these imperatives.

Despite the emerging consensus around the importance of teacher effectiveness among policymakers, there is a dearth of data to support the design and implementation of systems that can measure and support the effectiveness of classroom teachers. Amid these challenges, how can districts collect the right sort of data to measure teacher performance and drive systemic improvement? What legislative models enable districts to implement systems for evaluation that promote genuine and powerful professional learning? How can technology ease the financial and operational burden on district and school leaders implementing new teacher evaluation systems, while ensuring fidelity to implementation? Who will observe teachers, and how will we prepare them? How can districts ensure the accuracy of their judgments of teacher effectiveness?

In 2009, an ambitious effort was launched with a $45 million commitment from the Bill & Melinda Gates Foundation to "help educators and policymakers identify and support good teaching by improving the quality of information available about teacher practice."[1] Since 2009, the Measures of Effective Teaching (MET) project has enrolled approximately 3,000 teachers from the following seven school districts: New York, New York; Hillsborough, Florida; Memphis, Tennessee; Charlotte-Mecklenburg, North Carolina; Pittsburgh, Pennsylvania; Denver, Colorado; and Dallas, Texas. The MET project was designed to identify the combination of measures that work together to form a more complete indication of a teacher's impact on student achievement.

Teachscape served as the commercial contractor with responsibility for video-taping and scoring approximately 23,000 lessons, which served as the primary data set for the MET project. These 23,000 lessons were collected from more than 3,000 classrooms in grades 4–9 English language arts, grades 4–9 math, and grade 10 biology. The MET project enabled Teachscape to develop a unique perspective that marries a boots-on-the ground understanding of the challenges associated with implementation with a keen appreciation of the goals and needs of policymakers, teachers, and school leaders. After completing the video capture of lessons, there came the challenging task of scoring the classroom videos relative to multiple instructional frameworks. For this phase of the MET project, Teachscape partnered with the Educational Testing Service (ETS) to develop the systems, tools, and scoring methodology needed to complete this task. Dr. Catherine McClellan oversaw the MET scoring effort for ETS and was a major contributor to this paper.

The project concluded in the autumn of 2011 and its findings will be released in 2012. While much will be written about the MET research and its impact on teaching and learning, less has been shared about the tools, technologies, and processes that enabled its implementation and helped to garner support among teachers and school leaders.

The goal of Teachscape's practitioner series is to share its learnings from the MET project, in order to support district leaders and state-level policymakers in the implementation of new, impactful teacher evaluation systems.

This first paper addresses the often overlooked and complex set of challenges involved in training and certifying the evaluators of classroom teaching.

[1] See http://www.metproject.org/welcome

Executive Summary

In its October 31, 2011, Issue Brief entitled *Preparing Principals to Evaluate Teachers,* the National Governors Association made an urgent appeal to the nation's governors to take immediate steps "to ensure principals have the time they need to adequately train, become certified, and practice conducting evaluations *before* evaluation results are used to make high-stakes decisions."[2]

This paper recommends the following considerations be given to the design and implementation of programs to train and certify principals to conduct high-stakes teacher evaluations:

- Training programs must prepare principals and other classroom observers to understand the difference between bias, interpretation, and evidence.

- It is necessary but insufficient to teach observers the design and instructional philosophy behind the classroom observation instrument they will use to make high-stakes decisions about classroom teachers. The training must also require observers to accurately apply their knowledge of the instrument and demonstrate their ability to accurately score lessons from the range of grade levels and subjects that they will ultimately be expected to evaluate.

- An essential component of any training program is exemplar videos of classroom lessons that have been pre-scored by certified instrument experts, if not by the instrument's author.

- Because all classroom observation instruments are high-inference assessments, it is best to have more than one video illustrating "benchmark" performance on each score point on the rubric associated with the observation instrument. It is also important to have high and low "rangefinder" videos, in order to make clear to the trainee what the difference might be between a score at the high end of one performance level and a score at the low end of the next performance level on a particular rubric.

- There is no better training than authentic scoring practice. Whether using software or live classroom teaching with experts, good observer training will provide the opportunity to score authentic lessons and receive instant feedback from experts on the "true" scores for those lessons, along with explanations as to why the trainee's scores were correct or incorrect.

- Certification tests should assess the ability of the observer to replicate the scores of the instrument experts when observing a range of lessons in various grade/subject combinations.

- Certification tests should not only assess the ability of the observer to score accurately; they should also test the ability of the observer to get the right score for the right reason. This means observers must have the proper observation skills to collect all of the evidence from classroom practice that is relevant to each component of the scoring rubric they will use.

- Certification tests must assess the ability of the observer to differentiate between bias, interpretation, and evidence.

These recommendations and other findings emerged from Teachscape's work with Educational Testing Service on the Measures of Effective Teaching project, funded by the Bill & Melinda Gates Foundation.

[2] NGA Center for Best Practices, *Issue Brief: Preparing Principals to Evaluate Teachers,* October 31, 2011, p. 1.

Training Classroom Observers

As the evaluation of teachers is used for increasingly high-stakes personnel decisions, it becomes essential that the judgments made by evaluators are accurate and defensible, both professionally and legally. Until recently, the burden on teacher evaluators was modest: a cursory observation of teaching practice and the completion of a rudimentary teaching checklist were all that was required to determine that a teacher was performing at a high level.

But the policy landscape has changed dramatically. With the recognition of the vital role that teachers play in promoting student learning, it has become essential for evaluators to demonstrate that they can accurately assess (and diagnose for the purpose of supporting improvement) the quality of classroom instruction that they observe. And if an evaluator's judgment can have adverse consequences for a teacher's career, the importance placed on the accuracy of those judgments only increases.

To develop skilled classroom observers, training must be thorough, careful, and well structured. Observers' understanding of the application of the rubric must be reviewed frequently, and feedback that corrects misunderstandings must be given as soon as possible.

Content of Training

There are some components of observer training that are required, no matter what rubric the trainees are learning to use.

ORIENTATION

An orientation to the process of observation and the uses of the data from this process helps the trainees understand the importance of their attention and effort. If the training is online, an orientation to the software is essential so that unfamiliar tools do not detract from learning.

CONTROLLING FOR BIAS

Observer bias is an important issue that needs to be addressed directly in training. There are two types of bias that are especially important to address: bias due to observer preferences and bias due to observer knowledge of the candidate.

- Bias due to observer tendencies is one preference that thorough training can be expected to remove or at least minimize. It is important that this part of the training gets the trainee to recognize that everyone has biases and preferences. These can be based on a multitude of things: hair length, accent, teaching style, lesson content, clothing—even the paint color and arrangement of the classroom. The goal of bias training is not to remove these completely, as that is unrealistic, but to make trainees aware of their biases and preferences. Then they will be conscious of the consequences for scoring that can result and will make an effort to control them while scoring.

- Bias due to observer knowledge of the candidate being observed is far more difficult to address and control. It is virtually impossible to be completely objective and neutral when observing someone who is a long-term colleague or friend. Prior experiences with the candidate color the observer's perceptions in ways that are both obvious and subtle. Knowledge that the outcome data from the observation may have significant consequences for the candidate likely will incline the observer to be more lenient than is warranted by the strict factual evidence from the session. This component of observer bias is expected to be quite variable in size and direction, depending on the previous experience of the observer with the candidate.

UNDERSTANDING THE OBSERVATION INSTRUMENT

The first thing observers should learn about any observation instrument is basic background information. This includes how the instrument was developed, the pedagogical viewpoint or theory exemplified, a general description of the scoring scale and levels, and a list and brief description of the scales[3] the observers will be working with in their evaluation sessions. It is also important that the observer learn the proper protocol for conducting an observation. For example, it is imperative that observers learn how to collect a comprehensive set of evidence, and then to associate that evidence with the proper components of the observation instrument, *before* attempting to score the lesson. If different types of observations are required (entire class, short, or targeted to certain skills), a short summary of those should be included as well.

APPLYING THE RUBRICS TO OBSERVATION

The core content of the instrument will be covered in these sections. It is essential that these sections be delivered clearly, with careful thought for the construction and sequencing of information. Observers need to learn skills they will use in high-stakes situations here—it must be done right!

Observation rubrics have a set of ordered levels to which the observer assigns the skill level observed in the lesson (or portion of a lesson) on each scale. The levels can be assigned numeric scores (1 to 5, for example) or text labels (such as "basic," "effective," or "exemplary"); in either case, there is an ordering from least desirable to most desirable in the levels. On some rubrics, one or more scales may have reverse coding—where the low score is the most desirable and the high score the least. These negative scales are used often in instances such as a negative or hostile classroom climate, where a low score (implying few or no incidents) is "good" and a high score (indicating frequent or severe occurrences) is "bad." Reverse coding can be confusing for observers, and special attention should be paid to the training and resulting data from such scales.

Training typically proceeds through a consistent process: the trainees read the scale definition and an explanation and discussion of the definition; they study the different points on the scale and learn to differentiate the levels from one another. They then view videos of classroom practice that illustrate the different score points and acquire the skills of determining the level of performance represented in the video.

During training, the trainees should be expected to begin using the tools—including software—that they will use in live observations, beginning the process of familiarization as soon as possible (see "Process Considerations for Live Observation" below). While it is common to begin with the lowest or highest scale level and proceed through the levels, it may be best to start in the middle and work out to the low and then high cases. Most trainees can recognize extreme performances relatively quickly; the middle categories tend to be the most difficult to learn to distinguish, so experts recommend starting there.

Embedding quick knowledge checks into the training provides assurance that the trainees are indeed grasping the content. It also focuses trainee attention on the details when they realize that they may be quizzed. Even low-level, factual knowledge questions help reinforce the learning that must be in place to support the application of the rubric to an observation. If the training is online, such questions can easily be embedded into the flow of training, with written feedback for responses. In face-to-face training, while not impossible, it is more difficult to check each trainee individually.

After viewing the videos of classroom practice (see descriptions below) and internalizing the rationales for the scores given by experts, trainees should practice their new skills by scoring short videos. The more practice, the better! Each scale's training should conclude with a set of training videos on which the trainee scores and receives feedback. For every scale after the first one, the trainee should receive one set of practice videos to score on that scale only, and another set of videos to score on all scales learned

[3] Different observation instruments use the terms *domains, dimensions, components,* or others to refer to the units against which scores are recorded during an observation. For convenience and consistency herein, we will use the term *scale* to refer to this unit.

to that point in the training. This helps trainees build skills cumulatively, and will make the transition to using the full rubric at the end of training smooth and easy. It will also expose misunderstandings sooner. For example, a trainee who can score one scale in isolation but cannot score more than one scale simultaneously while maintaining accuracy and recording evidence should not continue further into the training. Instead, such a trainee should retrain on the current set of scales, until he/she is comfortable using the complete set of skills, before proceeding.

EXEMPLAR VIDEOS

Master-coded video exemplars[4] and rationales must be available in order to develop high-quality training. The video clips used in training are the means by which the abstract descriptions of the rubric take on concrete meaning. Training will not be successful without exemplar videos. Note that, for training purposes, it is not necessary, and can be detrimental, to show a trainee a full-length classroom lesson as an exemplar. Shorter clips, carefully selected so that the behavior or skill demonstrated is complete, are more effective. Trainee attention should be focused on the specific skill being taught.

A strong training program should include two types of exemplar videos—benchmark videos and rangefinder videos.

Benchmark Videos—A benchmark is a clear, unambiguous example of a scale level. It is in the middle of the scale level, not near the high or the low boundary. It is preferable that there be two benchmark videos for every scale level, for each scale in the training. A rubric of 6 scales with 4 levels each would require (6 * 4 * 2) 48 benchmark videos. The benchmark videos should be as different from each other as is possible while still remaining clear exemplars of the skill and level being shown. For each benchmark video, there should be an associated rationale, citing specific evidence in the video clip and tying it to the scale level descriptor, to explain why the video clip illustrates the scale level.

Rangefinder Videos—A rangefinder is a boundary case. These video clips should show cases that are a “low 2” or a “high proficient.” These exemplar videos are particularly important in that they assist trainees in learning to distinguish where the borders of the scale level are and on which side a particular case should be assigned. For each scale level, there should be a minimum of one low and one high rangefinder video; the exceptions are the lowest scale category, where a low rangefinder is not possible, and the highest scale category, where a high rangefinder is not necessary. A rubric that has 6 scales with 4 levels each would require (6 * 4 * [1 high + 1 low] – 2) 46 rangefinder videos. For each rangefinder video, there should be an associated rationale, citing specific evidence in the video clip and linking it to the scale level descriptor, to explain why the video clip illustrates the scale level. For each rangefinder, it is also important that the rationale include the reasons why the video clip is not classified in the adjacent scale level (e.g., explaining why a high 2 rangefinder is not a low 3).

Scoring Practice

As part of the training experience, trainees must score classroom lessons. The practice may be structured by the trainer or software, or done independently. The trainee should watch a video of teaching practice, capture evidence, and assign scale levels just as in a “real” observation. When complete, the correct scores will be revealed with the rationale and evidence explaining the score assignment. The feedback provided through the rationale helps trainees refine their skills in applying the rubric consistently and accurately. The scoring feedback provided should be specific to the score assigned by the trainee. If a trainee’s score is too low, the feedback should concentrate on evidence that may have been missed that matches the rubric at a higher scale level than the one assigned, and similarly for a score that is too high. Even for a correct score, the trainee should receive feedback that highlights the specific evidence that links to the correct scale level, in case the correct score was a lucky guess.

Practice video sets should require that the trainee use the full set of scales and should approximate the length of the observations the trainee will be expected to do once certified. If a trainee would be expected to observe a 45-minute class session, the trainee should practice

[4] Here “master-coded video exemplars” refers to videos that have been scored and annotated by experts in the instrument, typically the author(s) of the instrument itself. The process for selecting and adjudicating accurate “master scores” is as important as the selection of the master coders.

on sets of videos that are 30 minutes or longer to ensure that he/she can maintain attention and accurate evidence recording for the entire required period. There should be multiple practice sets, each with a range of performance requiring that the full range of scores be assigned. In these sets, it is also desirable that different content, grade level, and contexts (special education, ELL, lab settings, etc.) be included, assuring that the skills learned in training can be applied appropriately regardless of variation in such factors. Five practice sets of 10 videos would require an additional 50 videos with master codes and rationales being supplied.

TARGETED TRAINING SETS

If there are scales or levels that tend to be particularly difficult for trainees to master, it may be worth the investment required to create targeted training sets. For example, if a scale measuring formative assessment is challenging to learn because differentiating between "formative assessment" and "using questioning to deepen understanding" proves confusing, a set of practice videos with examples of one or both purposes in it could be created. While scored on all scales of the rubric for comprehensive practice, the emphasis in the rationales and feedback would be on detailed explanations of the difference between these two scales. A trainee having difficulty with this scale could be directed to complete this particular training set of videos to clarify misunderstanding. If trainees seem to have difficulty determining if a score level 1 or 2 should be assigned across all scales, a practice set of videos with performance concentrated in these categories could be created. Not every video should have scores of only 1 or 2 on every scale, as that would make the scoring of the set too easy—but the scale categories in question should occur more frequently than the others. If trainees are having difficulty with this scale level distinction, they could be assigned to practice on this set of videos.

Length of Training

Many people are surprised at how much time rigorous training for observers requires. Observers are being trained to do an activity that requires they watch a classroom session that is between a half-hour and an hour. The session is full of complex interactions between teachers and students. In order to learn the skills, it is necessary that trainees watch numerous examples of these interactions and make the connections with how the rubric evaluates the evidence seen. Trainees must learn to put aside personal preferences that have developed over years or lifetimes. They must learn to record evidence using novel tools and systems. Each trainee must learn to see accurately and consistently through the lens of a complex rubric. This skills system takes time to construct.

GUIDELINES ON THE LENGTH OF TRAINING

- Introduction, bias, process and tools: 3–4 hours

- Scale training, including content definitions, number of levels and descriptions, viewing benchmarks and rangefinders: approximately 30 minutes—1 hour per scale level, assuming 2 benchmarks and 2 rangefinders of approximately 8 minutes each. Note that this implies 20 minutes of video viewing time for each scale level; reading the definitions, collecting evidence, and assigning scale levels require additional time, as does reading rationales and viewing videos again for evidence missed. A rubric with 6 scales and 4 levels each would require between 12–24 hours of content training.

- Practice scoring embedded in training: 2 hours per scale, assuming 10 videos (2 sets of 5 clips each) of approximately 5 minutes for each scale, plus scoring time and review of rationales. A rubric with 6 scales would take about 12 hours for practice scoring embedded in the training.

- Full-length practice sets: 8–10 hours per practice set, assuming 10 videos of at least 30 minutes each. Note that this requires a minimum of 5 hours of video viewing; evidence collection and assignment of scale levels require additional time, as does understanding rationales and viewing videos again for evidence missed.

Assuming a rubric of 6 scales and 4 levels each, training and practice could be expected to take between 35 and 50 hours if trainees complete one practice set of videos. Obviously, the time required will vary depending on the complexity of the rubric, the length of the exemplar video clips, and the amount of practice included in training.

Training Format: Face-to-Face or Online?

There are advantages and disadvantages to both face-to-face and online training. Many trainees express an initial preference for face-to-face training. Most enjoy the social aspects and the ability to ask questions in this format. Trainers also enjoy the social aspects of face-to-face, and they like the ability to monitor the expressions and body language of the trainees; however, face-to-face training is not without drawbacks. Even with the best trainers, there is, inevitably, individual variation in how the training is delivered, both across different trainers and even with the same trainer on different occasions. This can result in different interpretations of the content by the trainees. The social aspects of face-to-face training may discourage sustained, focused individual learning activities, an essential component of high-quality observer training. Furthermore, face-to-face training usually requires the trainees to attend at some central location at specific times when the trainer is available. For some trainees, the requirement for travel can be prohibitive and burdensome. If the training extends beyond a single day, the costs of travel, housing, and subsistence for the trainees—in addition to the meeting space—must be considered.

Online training, although initially less popular, has been shown in prior research to be equally effective to face-to-face. This has been shown in a wide variety of fields, including writing assessment (Wolfe, Matthews, & Vickers, 2010; Zhang, Powers, Wright, & Morgan, 2003), history marking (Chamberlain & Taylor, 2011), second-language essay scoring (Elder, Barkhuizen, Knoch, & von Randow, 2007; Knoch, Read, & von Randow, 2007), emergency medical team training (Heinrichs, Youngblood, Harter, & Dev, 2008), and psychiatry (Kobak, Englehardt, & Lipsitz, 2005), among others. The loss of direct interaction means that structures such as FAQs, webinars, and direct interactive feedback must be created to respond to trainee questions. The anonymity of electronic communication can encourage less confident individuals to explore concepts that they did not previously understand; something they may not do in a face-to-face setting. One advantage of online training is its absolute consistency: the same material is delivered in precisely the same way to all trainees every time.

The material can be reviewed at any time with any frequency by each trainee, without frustrating the trainer or the other trainees. Online training generally can be viewed at any time of day or night, at the convenience of the trainee, and no travel is required. While online training can be difficult and costly to develop, it is, arguably, a better approach for training observers.

Professional Development for Teachers

Fairness (and potentially the collective bargaining agreement) dictates that those being evaluated should know the evaluative criteria on which their performance will be evaluated. In the case of teacher observations, this implies some level of training for the observed as well as the observers. Training for teachers who are to be observed does not need to be as extensive as that for the observers, and they do not need the skills assessments and practice the observers do. However, experiencing the core content training will increase teachers' confidence in the observation system, and because teachers become more conscious of the behaviors that the rubric considers desirable and effective, improved practice is often an attractive byproduct of this training.

If training is built online, a subset of the full observer training could be selected for teachers. If face-to-face, a separate session would need to be conducted. The most important sections are the core content training (described above). The other section to consider for inclusion for teachers would be the introduction, bias training, and observation process and tools. Seeing the rubric, the complexity of the judgments, the evidence required, the standards to which the observers will be held, and what tools and procedures will be used both demystifies the process and increases confidence that the process is both valid and fair. Feedback from an observation will make more sense if the observed teacher has been trained in the rubric scales and levels and knows

Certification of Classroom Observers

Those who mandate high-stakes evaluations of teachers based in meaningful part on classroom observations have an ethical and, potentially, a legal obligation to verify the skills of the observers charged with conducting those high-stakes observations. If the observer is not demonstrably accurate, fair, impartial, and consistent in scoring observations according to the rubric, the judgments made as a consequence of the observation will be open to challenge. At the end of training, observers should be assessed to verify that they have learned the information presented in the training, and that they can apply the rubric accurately and consistently. Such an assessment will be referred to herein as an observer "certification" test. There are a number of considerations that go into developing a certification test that is defensible.

PURPOSE

As noted, the main purpose of the certification test is to separate observers who can demonstrate an acceptable level of skill in applying the observation rubric from those who cannot. While this sounds simple enough, many who have tried to develop a high-stakes assessment have discovered that it is not easy.

One way to think about certification test construction is to consider what claim(s) the test results must support: What do you want to be able to say with confidence about the observer who passes certification; that an observer can apply the observation rubric accurately and consistently? Fine, but in what context? When observing a teacher at any grade level (pre-K through grade 12) to the teaching of any content matter (calculus *and* studio art *and* PE *and* early reading skills)? Observing any population of students (ELL, for example)? If you want to claim all of those things, then you must assess all of those things—or at least have a strong basis for generalization to those things from the specific aspects you do assess.

RELIABILITY

It is common to hear "rater reliability" discussed in the context of teacher observation. In most cases, the discussion is about inter-rater agreement: the extent to which two independent observers assign the same score or set of scores to the same classroom session. As important as it is, rater reliability is a topic for another paper. The aspect of reliability important in observer certification is the extent to which the outcome of an assessment would remain the same if the observer were tested again (with no memory of the first assessment) on the same skills, or were tested with an alternate but parallel form of the test. Would we make the same decision about whether an observer is acceptably accurate and precise if we tested that observer with an alternate test form or on an alternate occasion? This type of reliability is a property of the scores on a measure, not of the measure itself, and depends on the population of examinees. There are a number of statistics used to calculate this type of reliability. It is worth noting that test reliability is expected to be quite high—often 0.80 or higher—when the results are used to make consequential personnel decisions.

VALIDITY

Here we refer not to the validity of the observation rubric (also a topic for another paper), but the validity of the observer certification assessment. Validity, in simple terms, is the degree to which the interpretation and use of the results of an assessment are supported by evidence and theory. Note some important facts about validity:

- Validity is not a property of a test or a score—neither a test nor a score can "be valid." Instead, it is a property of the *use* of the test or score. Validity is contextual. The use and interpretation of the scores from a test can and should be valid, as can decisions or actions based on those results.

- In classical test theory, a test can be no more valid than it is reliable; reliability bounds validity. A test can be reliable but not be valid for a particular use; the reverse (that a test can be valid and not reliable) cannot occur. A test that cannot be used to measure the same thing consistently cannot then produce results that can be interpreted or used to make valid decisions or choose valid courses of action.

- As a simple example, consider measuring the height of an adult. If the scale of measurement you are given to do so is marked in units of miles, reliable measurement will be difficult. Different people using the scale to measure the same person will arrive at different values—the measurement is not reliable. However, given the same task and a measuring tape marked in inches, different people probably will report nearly the same value—their measurement is reliable. The second set of measurements would support a valid inference about the height of the person measured, within the precision of the measurements taken. It would not, however, support a valid inference about how much calculus that same person knows. For that purpose, even though the measurement of height is reliable, it is not valid.

Establishing the validity of the use of scores from a measure is not a trivial undertaking. Validity can be divided into numerous subtypes, each with requirements necessary to support claims for it. Some require expert review of content; some require data analysis; some require the use of criteria external to the assessment system. All require thorough documentation and independent review to ensure that the work meets the standards of the field.

For an observer certification assessment, each observer must demonstrate the predefined minimum level of skill, and that minimum level must be equivalent as measured on different forms of the assessment if there is more than one form, or on different occasions if an examinee is assessed more than once.

Assessment Design

There is a long history of practice and science in designing and developing assessments. Testing companies have large numbers of professionals and highly evolved processes in place to do this. The result of the investment made up front yields benefits in highly reliable tests that produce scores that are shown to be valid to inform decisions and actions. For many professionally developed, high-stakes exams, the uses and interpretations of scores have been challenged—and upheld—in courts of law. There is every reason to believe that the process by which a school district conducts high-stakes observations of teaching will ultimately be adjudicated by the courts as well.

ASSESSMENT ITEM TYPES

Properties of different types and formats of assessment items have been researched extensively, and are chosen to balance constraints of cost, time, and examinee burden while meeting requirements as to content or construct coverage and difficulty. The two large classes of item types are selected response and constructed response. Selected response items require the examinee to choose a response from a set provided by the instrument; for example, multiple-choice or matching. Constructed response items require the examinee to create a response independently based on the stem or prompt given on the instrument; examples of these item formats include essays written on an assigned topic, art pieces, musical performances, and spoken responses. Constructed response items generally are scored by humans using a rubric created as part of the item; the rubric defines the scales (if more than one) on which the response will be judged, sets the number of score levels for each scale, and describes the characteristics of the responses that belong in each score level.

Observations of teaching are a special case of constructed response items. Usually, there is no prompt or stem: teachers are expected to conduct whatever lesson was planned for the class session that day, regardless of the presence or absence of an observer. There may be special preparation or materials required, such a detailed lesson plan, a pre-observation description of the goals of the lesson, or a post-observation reflection on what happened in the lesson, both good and bad. As with other constructed response items, the observation must be scored by a human being using a rubric, where the rubric defines a set of scales, with levels and descriptions for each scale.

Since an observer certification test is intended to assess skills in classroom observation, logic would seem to dictate that the assessment should require the observer to score class sessions. This would be ideal in many ways. Let's think this through using an example. Assume that an observer to be certified will observe all teachers in his or her elementary school. The (oversimplified) school is structured as follows:

- Grades 1–4

- Subjects taught at all grades are math, reading, writing, social studies, science, PE, art, health, and music. There are focused classes for students with reading disabilities and with cognitive disabilities (mild and profound) as well as English language learners.

- The classroom teachers at each grade level teach math, reading, writing, social studies, and science; there is one specialist teacher for each of the other subjects.

An observer evaluating teachers in this school must be able to assess performance in 4 grade levels by 9 content areas or specializations at each grade level, which equals 36 combinations of grade/content. Aside from all other considerations (and we'll come back to those), it is poor practice to assess anything with a single item. So with only 2 items per grade/subject combination, our observer certification test now has 72 items. That doesn't sound too bad until you realize what an "item" in this context is: an observation of a class session, or in the testing context, a classroom session video. These videos can be 30–60 minutes long, so our test now requires the examinee watch between 36 and 72 hours of classroom video—ignoring any time required to take down evidence or score! Allowing evidence and scoring time, this test likely would be between 54 and 108 hours long, far too long for an examinee to realistically be expected to sit. So how can that time be reduced?

Remember that tests are designed to balance cost, time, and examinee burden while meeting requirements as to content or construct coverage and difficulty. Reflecting back to the discussion of item types, constructed response and selected response items have different properties and make different demands on examinees. Despite a reputation for assessing only memory and low-level skills, well-designed selected response items can assess complex traits and higher-order thinking and problem-solving skills. Selected response items confer advantages in terms of test reliability and in time required, so more selected response items can be administered reliably in the same amount of time.

Test Item Types

It seems that including some selected response items on an observer certification test is a good idea for reducing the time and improving the test reliability, but what kind of questions should be asked that fit the construct? The connection with classroom observation, rubric knowledge, and accurate application must remain strong. The following are some suggested item types:

- Including assessments of knowledge of the observation rubric, as long as that is *not* the only thing assessed on a certification test, is perfectly appropriate. It allows examinees to demonstrate that they have been attending to the training and have learned the basic information about the instrument they will be using.

- Using short video clips from classroom sessions will reduce the time required for the test. Such clips must be chosen so that the clip illustrates a particular type of behavior or scale level occurring on the rubric. Short video clips should not be the entirety of the observations completed on the certification test, as they are not a realistic parallel to the situation in which the skills will be used, but they can provide useful insight into the ability to locate and describe evidence associated with particular scales and levels. The selected response format permits assessment not only of the observer's accuracy in assigning the correct scale level, but also of skills such as an observer's ability to discriminate between appropriate and inappropriate evidence to support a particular scale level assignment, something difficult to do in another context or format.

- Observers should be assessed with some number of full-length class session videos, since these most closely parallel the setting in which the teaching observations will occur. Examinees should watch and score these videos just as they would a "live" observation, using the tools, language, and standards of the rubric. Selected response items can be administered after the examinee has watched the video, captured evidence, and assigned scale levels. For example, examinees could be asked to match selected evidence with the approximate time it occurred during the video session using only the examinees' notes and evidence. A selected response item could ask the examinee to watch a specific time segment and choose the most important event relevant to a particular scale during that period.

If the item formats are chosen to maximize the use of the information available in the videos, a reasonable balance of cost, time, and examinee burden can be created while still thoroughly assessing the skills and knowledge of the observer.

Passing Standard

The performance standard required to pass an observer certification test should be developed with care. Given that high-stakes personnel decisions will be made using the data from the observations, the standard required should be quite robust. Observers must be accurate and consistent in applying the rubric and be able to demonstrate this at a high level. The optimal way to determine the passing standard on a certification test is to conduct a formal standard-setting study. There are a number of standard-setting procedures often used in educational testing, and there is extensive literature describing them. However the passing standard is set, if it is to be defensible, it cannot be arbitrary. While setting passing standards necessarily has a judgmental component, the process of determining a passing standard should be a systematic approach to translating expert judgment of minimal competence into an operational point on the test's reporting scale.

Assessment Administration Considerations

There are a number of administration concerns that should be attended to as part of the test design.

SECURITY

One aspect that receives less attention than warranted is test security. We have referred to the observer certification test as high stakes more than once. Is it really? The observation of teaching using complex rubrics is becoming an important component of school administrators' responsibilities and thus must be part of their performance evaluation. Being certified to observe and judge teaching using a rubric will be a required job task; lacking the certification, the administrator can no longer do his/her job. It is not unrealistic to imagine that a lack of certification to complete observations of teaching would result in action up to and including termination. This is high stakes, therefore, not only to the teachers being assessed, but to the administrators conducting the evaluations!

Unfortunately, high-stakes tests bring with them requirements for high levels of test security. While one would hope that no one would cheat on a certification test, recent news stories have shown that intense pressure can lead to aberrant behavior. Breaches of test security carry a cost to everyone: those attempting to cheat who face disciplinary measures; the examinees who fairly earned the credential tarred unfairly with the same brush; and the test administrator who must create new forms and items with resources, money, and time better spent elsewhere.

One measure of test security that should be considered very seriously is proctoring the certification test by an independent party with no stake in the system. If the test is administered online, it would be preferable to have the examinees assessed on a computer that is not their own and one that they do not have access to after the test is over, so that items or answers from the certification test cannot be copied and subsequently shared. (A school's computer lab could meet most of these specifications.) All examinees should be required to present identification before taking the test, and to sign an honor pledge that the answers they give are their own and that they have received no outside assistance.

There are a number of approaches to detecting cheating on tests, such as the comparison of incorrect responses, erasure analysis (for paper-and-pencil tests), and similarity detection analyses for constructed responses. Such methods should be used on certification assessments, as they would be on any other high-stakes test.

TEST FORMS

Another method for increasing test security is the creation and use of alternate forms. It does an examinee who intends to cheat no good to have the answers to Form A if he/she is given Form B when tested. Creating multiple forms requires an investment of resources, however. Multiple forms of the "same" test must be as parallel as possible in all aspects: content, difficulty, format, and item types. Observer certification tests are not simple to create and require a large amount of master-coded video. The more test forms that are required, the more work must go into the selection, master-coding, and rationale-writing of videos and clips, as well as item development. In addition to increasing security, examinees often are permitted more than one attempt at certification if they fail the first attempt. If two attempts are permitted, then two forms are required; three attempts require three forms, and so on. If the forms are believed to be genuinely parallel and interchangeable, randomly selecting the form to be administered for each examinee reduces the probability of successful cheating.

An advantage of online administration is that software systems can be programmed to create test forms from pools of items, as long as the test meets certain constraints. The test forms can be entirely fixed *a priori* to entirely computer adaptive, at the extremes. In one basic version, the items for a test form are fixed, but the order is scrambled for each examinee. As long as there is no reason to expect context effects from the order of item presentation (and this is an important condition!), scrambling should reduce the probability of successful cheating. Since fully computer-adaptive systems have proven expensive in terms of item exposure and pool maintenance, they are not attractive choices for an observer certification test. More likely in this context is a system that can do linear on-the-fly test (LOFT) construction. In this design, a set of statistical and content specifications is established, and the software pseudo-randomly selects items within those constraints so that each examinee receives a (potentially unique) test that meets the specifications.

Frequency of Recertification

Certification provides assurance that, at the time the assessment was completed, the observer knew the required information and could apply the rubric at the required level of accuracy and consistency. It does not and cannot guarantee that this observer will always demonstrate this level of performance. People forget, and skills not in continuous use tend to diminish. Even skills in regular use can shift given the context of use. In planning a teaching observation system, it is important to keep these facts in mind when considering the issue of how long an observer's proficiency is in force.

To address this issue of observer drift, a set of tools can be implemented to locate and correct factors. Tools used inside the observation system, such as calibration, double-scoring, and validity scoring, will be addressed in another paper. In addition to those tools, regular recertification of observers should be required; preferably in tandem with some level of review of training materials. This should occur annually, at a minimum, and logically would be scheduled at the end of the summer break. Since an observer is unlikely to have performed many, if any, observations during summer, the skills review and recertification are both most likely to be needed and to be most valuable at that point. It is also reasonable to believe that observers would have the necessary time to complete the review and assessment just prior to school starting in the fall.

Process Considerations for Live Observation

In order to maximize the value and quality of the data gathered from an observation session, it is essential that the tools used to collect data not interfere with the "work" of observing. Observers must be familiar and comfortable with any paper forms, software, video and audio recording devices, data entry tools, and the style and content of evidence capture from each observation session. This can be achieved through regular and extensive practice in using the tools of data collection in situations similar to the live observation, until their use is automatic.

Like any form of judgment, accurate classroom observation requires observers to base their judgments on the preponderance of the evidence. Most observational instruments require the observer to take extensive notes (and sometimes collect classroom artifacts) relevant to the teaching, learning, and interactions they observe in the classroom. There are distinct skills associated with accurate evidence collection. Observers must be trained to "see" what is transpiring before their eyes through the lens of the specific observation protocol and the evidence the rubric deems significant in classification. Does the teacher interact differently with male or female students? How many different students were engaged throughout the lesson? How many times, or how frequently, did the teacher check for understanding? When correcting a student's misperception of an algorithm, did the teacher get the math right? What depth of questions were asked and answered, and by whom? Did the students or the teacher dominate the discussion?

Video-based observations make it easier for observers to see what actually transpired during the lesson, with options to pause or rewind and watch events again. In a live classroom observation the accuracy of an observer's score, and the observer's ability to ultimately defend that score if challenged, is based entirely upon the depth and accuracy of the evidence collected during the observation. High-quality observer training provides extensive opportunities for trainees to compare what they took down as evidence, see what master observers captured, and reduce or eliminate the differences.

Continuous Improvement and Support for Observers

Observers who have been trained and certified have reason to be proud of their accomplishments. They have mastered a difficult task to a high standard of performance. However, there is always room for improvement. Because the task is complex, observers can continuously make small improvements in their application of the rubric, their skills in collecting evidence and assigning it to the appropriate scale, their ability to follow and dissect complex classroom interactions, and their ability to translate an accurate scale level assignment into useful and actionable feedback.

In order to achieve this type of improvement, support must be in place. Opportunities to discuss observations with colleagues who also use the same rubric must be provided. Guidance from an expert in the rubric should be provided as part of professional performance feedback. There should be a forum for sharing questions, insights, and tips with others, both locally and beyond. Observers must practice in settings where they are a "neutral observer" as well as in their home school. As classroom observation becomes a key aspect of job performance, a system of rewards for continuous improvement in observation skills must be created as well.

Conclusion

Developing a system of observer training and certification is one crucial component of a complete teaching observation and teacher evaluation system. Whether the work is undertaken at the local level or a system is purchased from a vendor, the same high standards of practice will be expected. Challenges to highly consequential decisions are nearly certain to occur, at local boards as well as in the courts. In order to have valid, reliable, and defensible grounds for actions, the foundation upon which the actions are built must be strong.

The more transparent the system is to the teachers observed, the more confidence they will have in the outcomes. If an educational authority has selected or created an observation system that is well planned and executed to the standards of best practice, the results will be improvement in teaching, in student performance, and in satisfaction for all stakeholders.

References

Chamberlain, S., & Taylor, R. (2011). Online or face-to-face? An experimental study of examiner training. *British Journal of Educational Technology, 42*(4), 665–675.

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing, 24*(1), 37–64.

Heinrichs, W. L., Youngblood, P., Harter, P. M., & Dev, P. (2008). Simulation for team training and assessment: Case studies of online training with virtual worlds. *World Journal of Surgery, 32*(2), 161–170.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12*(1), 26–43.

Kobak, K. A., Englehardt, N., & Lipsitz, J. D. (2005). Enriched rater training using Internet based technologies: A comparison to traditional rater training in a multi-site depression trial. *Journal of Psychiatric Research, 40*(3), 192–199.

NGA Center for Best Practices. (2011). *Issue brief: Preparing principals to evaluate teachers.* Washington, DC: Author.

Wolfe, E. W., Matthews, S., & Vickers, D. (2010). The effectiveness and efficiency of distributed online, regional online, and regional face-to-face training for writing assessment raters. *Journal of Technology, Learning, and Assessment, 10*(1). Retrieved October 31, 2011, from http://www.jtla.org.

Zhang, Y., Powers, D. E., Wright, W., & Morgan, R. (2003). *Applying the online scoring network (OSN) to advanced placement program (AP®) tests.* [ETS RR-03-12]. Princeton, NJ: Educational Testing Service.

teachscape

877.988.3224
info@teachscape.com
71 Stevenson St., 5th Floor
San Francisco, CA 94105
www.teachscape.com