WestEd
WestEd.org

MATH in common™

Classroom Observations

Documenting Shifts in Instruction for Districtwide Improvement

Rebecca R. Perry
Nanette M. Seago
Elizabeth Burr
Marie Broek
Neal D. Finkelstein

January 26, 2015

FORMATIVE EVALUATION CYCLE REPORT FOR THE MATH IN COMMON INITIATIVE • VOLUME 2

2015 © WestEd. All rights reserved.

Suggested citation: Perry, R. R., Seago, N. M, Burr, E., Broek, E., & Finkelstein, N. D. (2015). *Classroom Observations: Documenting Shifts in Instruction for Districtwide Improvement.* San Francisco, CA: WestEd.

Requests for permission to reproduce any parts of this report should be directed to WestEd Publications Center, 730 Harrison Street, San Francisco, CA 94107-1242, 888-293-7833, fax 415-512-2024, permissions@WestEd.org, or http://www.WestEd.org/permissions.

Table of Contents

FIGURE

TABLE

WestEd's Evaluation of the Math in Common Initiative

Math in Common™ is a five-year initiative funded by the S. D. Bechtel, Jr. Foundation that supports a formal network of 10 California school districts as they are implementing the Common Core State Standards in mathematics (CCSS-M) across grades K–8. Math in Common grants have been awarded to the school districts of Dinuba, Elk Grove, Garden Grove, Long Beach, Oakland, Oceanside, Sacramento City, San Francisco, Sanger, and Santa Ana. California Education Partners provides technical assistance in support of the Math in Common Community of Practice. WestEd is providing developmental evaluation services over the course of the initiative. The evaluation plan is designed principally to provide relevant and timely information to help each of the Math in Common districts meet their implementation objectives.

The overall evaluation focuses on four central themes, which attempt to capture the major areas of work and focus in the districts as well as the primary indicators of change and growth. These themes are

- Shifts in teachers' instructional approaches and the corresponding teaching quality related to CCSS-M in grades K–8.
- Changes in students' proficiency in mathematics, measured against the CCSS-M.
- Change-management processes at the school district level, including district leadership, organizational design, and management systems that specifically support and/or maintain investments in CCSS-M implementation.
- The development and sustainability of the Math in Common Community of Practice.

Districts participating in the Math in Common initiative are diverse, ranging from small rural districts to large urban districts. Each district's unique context and history play a role in the path district educators will take in responding to the new instructional demands of the CCSS-M and in determining district-specific priorities regarding teacher professional development, aligned instructional materials, and assessment of student learning of the standards. However, participation of these diverse districts in this Math in Common Community of Practice also enables them to learn from each other through sharing their progress and successes, as well as their challenges and lessons learned. WestEd's evaluation activities will draw on the various district contexts to highlight how the districts, funder, and broader community can learn from the efforts of these 10 districts to implement the CCSS-M.

Introduction

Even seemingly straightforward education policy ideas are interpreted and implemented quite differently as they make their way through the levels of the education system (Cohen, 1990; Cohen & Hill, 2001; Spillane, 2000). Complex ideas that lack clear and specific instructional guidance, like the Common Core State Standards in mathematics (CCSS-M)—with their increased emphasis on rigorous and coherent content, standards for mathematical practice, and instructional pedagogies that support students' deep conceptual mathematics learning—may prove challenging as teachers attempt to interpret and implement them in their own classrooms. The combination of limited instructional guidance for the CCSS-M and individual teacher variation (resulting from each teacher's different beliefs, skills, knowledge, and interests) leaves room for significant variation in how the central CCSS-M reform ideas are interpreted and implemented in the classroom. As such, there will likely be wide variation in teachers' instruction as they implement the CCSS-M in their classrooms.

Yet if, as research has shown, teachers affect student achievement more than any other school-related factor (Rivkin, Hanushek, & Kain, 2005), Math in Common districts will need to understand and monitor how CCSS-M ideas are taught in classrooms in order to improve mathematics education for all students. Understanding the extent of teachers' instructional variation will help districts build on and spread best practices and support improvement of CCSS-M implementation.

IMPLEMENTING CLASSROOM OBSERVATION SYSTEMS

A classroom observation system (including both an observation tool and a protocol for its use) is one key support for documenting such instructional shifts resulting from implementation of the CCSS-M. However, developing valid and reliable systems for classroom observations is not easy. As a recent Carnegie Foundation report stated, "One common question still begs to be answered: what exactly are the most effective teachers doing that is working so well? If identifying the best teachers is complex and controversial, the process of identifying what they are doing promises to be even more so" (Stewart, 2006, p. 14).

Classroom observations have been used to study program delivery and policy implementation outcomes in education since federal education funding began in the 1960s (see, e.g., Rosenshine & Furst, 1973), yet interest in classroom observations grew substantially as a result of Obama's Race to the Top initiative. Although the majority of observation systems of this recent era aim to achieve the same goal of measuring teacher effectiveness, the observation systems that have been developed take many forms. The observation tools measured different aspects of teaching effectiveness, varied greatly in their implementation, and, as a result, were differentially useful for informing policy decisions and ongoing improvement efforts (Goe, Bell & Little, 2008).

With the recent demands of understanding and improving implementation of the CCSS-M, districts are again considering how to use classroom observation systems to document mathematics teachers' instructional shifts in relation to the standards and how to use such documentation to provide instructional feedback to teachers, allocate resources, and shape policy decisions about instructional materials. As the 10 California Math in Common (MiC) districts are each working to devise classroom observation systems to document such instructional shifts in order to inform districtwide improvement, certain patterns have emerged:

- Districts are organizing classroom observation systems somewhat differently, informed by their own local contexts and questions about instruction.
- Districts are finding that there are many difficult decisions to be made with respect to observation systems—for example, whether to use an existing instrument or develop a new one—and that putting in place high-quality observation systems requires significant planning and manpower to organize and implement.

To help districts working through these issues, WestEd has created this brief research synthesis and annotated bibliography to answer the question, "What does the recent literature say about selecting or developing and using classroom observation systems in general and more specifically in mathematics for the purpose of documenting instructional shifts?" This research brief provides Math in Common districts with research about the considerations necessary for selecting or developing and using classroom observation systems to document instructional shifts and to inform district improvement efforts. The report is neither intended to be a comprehensive review of the literature nor a prescription for district action, but a means of highlighting critical features of classroom observation systems and steps that will need to be taken in their development and use. Because the Math in Common districts have unique contexts and goals, each district will need to consider how to translate these theoretical ideas into their own best practices.

The report is organized into three main sections. First, we briefly explore what the research literature says about existing observation systems, and we highlight several design considerations for successful observation systems. Second, we discuss in detail several considerations of these findings for the Math in Common districts as they implement observation systems in order to better track and understand how teachers are implementing the CCSS-M in their classrooms. In the third section of the report, we provide an annotated bibliography for a number of recent publications on classroom observations that may be of interest to Math in Common district representatives interested in exploring these ideas in more depth. (Appendix A describes the method for inclusion of these resources.)

Central Design Features of Existing Observation Systems

There are a plethora of existing classroom observation tools and possible adaptations that can be made to them. Importantly, the research literature tells us that there is no one "right" observation system sufficient for all contexts and conditions; depending on local contexts and specific observation practices, different strategies might work better for some school systems than others (Jerald, 2012a; White, 2014). The literature does identify three main themes about designing successful classroom observation systems that should be considered as they are put in place: (1) purpose/use; (2) focus, and; (3) reliability of the observation data. We briefly describe the research findings on these themes prior to discussing implications for the Math in Common districts in implementing their own observation systems.

PURPOSE/USE

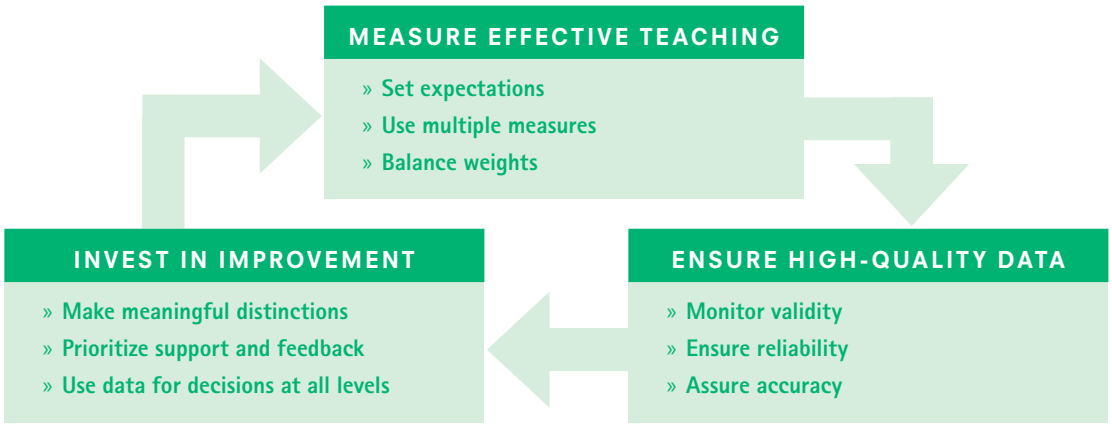
Because there can be multiple possible purposes for conducting classroom observations, "It is critically important to clarify observation goals at the outset of a project. The goals will determine how, when, and who you observe, and those decisions will influence how you can use the data you have collected" (Vitiello & Hadden, 2014, p. 13). The authors suggest that in order to clarify goals, it may be useful to consider the following questions:

- What do you hope to accomplish by conducting classroom observations?
- Who are the stakeholders for the observation data?
- Are you interested in the effectiveness of individual classrooms or in trends in instructional effectiveness more broadly?
- Are your goals more formative or summative in nature?

Others agree that it is important to design observation systems with "the end" in mind (McDonald Connor, 2013); that is, it is important to have a clear understanding of ways in which observation data may be used to explore relationships between policy decisions and practice.

Figure 1 may be useful to illustrate the cyclical relationship between classroom observations, high-quality

Figure 1. A framework for improvement-focused teacher evaluation systems



Source: The Measures of Effective Teaching Project, 2013

Does the observation system have the correct focus?

Districts can use the following questions (from Milanowski, Prince, & Koppich, 2007) to inform the focus for their observation tool.

1. Are the dimensions . . . that the system measures the drivers of important outcomes, such as student learning?
2. Are important drivers missing?
3. Does the system have so many dimensions that the key drivers get lost?
4. Does the system include a way to measure what truly distinguishes an outstanding performer from an average performer?

"[What] the system measures . . . should directly reflect what educators need to do to carry out the organization's strategies for achieving its goals" (2007, p. 3).

formative data gathering, and subsequent investments in improvement (The Measures of Effective Teaching Project (MET), 2013). While most classroom observation systems developed as part of the teacher effectiveness research previously focused on evaluation of and investments in improvement for *particular teachers*, much of the current research shifts the emphasis of classroom observations away from teacher *evaluation* and more toward teacher *feedback and support*—that is, using observation systems as "key levers for improvement of teaching" (Hill & Grossman, 2013). Because of this distinction, observation systems built for accountability purposes may not be dually suitable as systems for supporting teacher improvement; complementary observation systems for informing instructional improvement will likely need to also be put in place (Hill & Grossman, 2013; MET Project, 2013). Similarly, observation systems meant to accomplish other goals—such as to study the impact of allocated resources or the relationship between policy measures and practice—will likely need to be developed with these purposes in mind (Jerald, 2012a; Pianta & Hamre, 2013; Stuhlman, Hamre, Downer, & Pianta, 2014).

FOCUS

Determining a clear focus for the observation system is critical to ensure that it is relevant and useful for the district. No observation system can accomplish all goals, and by focusing on any one activity or aspect of instruction, others are likely to be lost (Harvey, 2006; Rosenshine, 1970). Prioritizing the critical features of the observation tool is important, and the narrowness or breadth (i.e., "grain size") of the observations should be dictated by the overall purpose (Hill & Grossman, 2013). The research literature generally provides two lessons on how to focus an observation tool:

- Focus the observation tool on the instructional shifts or other aspects of the learning environment you most want to understand.
- Choose a grain size that allows the observation system to be simple enough to use[1] and will result in data useful for its intended purpose.

The call-out box titled "Does the observation system have the correct focus?" provides some questions that might inform a district's focus for its observation system. For example, if a district presumed that teacher professional development was an important "driver" of student mathematics achievement, and wanted to use an observation system to document and inform professional development around teachers' use of multiple mathematical representations, the tool would need to focus on capturing the range of teacher and/or student actions that would

[1] "Simple enough to use" is determined by the focus of the observation system and by the features of how the observation system will be implemented, including decisions about whether to conduct live or video-recorded classroom observations, how much data to collect, and how the data will ultimately be used. We further address these implementation issues in the subsequent section on considerations for Math in Common districts.

demonstrate successful and less successful examples of using multiple mathematical representations.

An examination of several existing observation tools (see Table 1 on page 6) illustrates how wide-ranging observation foci can be. For example, the Classroom Assessment Scoring System (CLASS) is grounded in models of effective teaching and sets of teacher performance standards, and is generic with regard to subject matter and can be used across grade levels and content areas (Youngs, 2013). The CLASS captures a wide range of teacher behaviors (e.g., instructional support, classroom organization, emotional support). Other tools have different foci—for example, the Mathematical Quality of Instruction (MQI) tool (University of Michigan, 2006) aims to identify levels of "rich mathematics" instruction, and the TRU Math tool (Schoenfeld, 2013) captures aspects of teaching believed to be consequential for students' development of robust algebraic understandings. Still other observation tools, such as the Strategic Education Research Partnership's 5x8 card, are specifically intended to focus on student actions rather than teacher actions (New Teacher Project, 2011).

Marzano & Toth's (2014, p. 11) observation tool is focused on instructional shifts related to the CCSS and on students' activity during the lessons. The authors assert that this focus on instruction and student activity can result in information that is useful for supporting students to achieve rigorous academic standards. Using this observation tool, their research found that classroom instruction was most frequently devoted to introducing and practicing new knowledge rather than providing opportunities for students to engage in "cognitively complex tasks involving generating and testing hypotheses" that may support attainment of CCSS (Marzano & Toth, 2014).

The *validity* of the observation system is a critical feature to consider as part of the focus. Specifically, validity concerns whether the observation system is accurately measuring what it is designed to measure (e.g., specific instructional shifts, student learning; Milanowski et al., 2007; Sartain, Stoelinga, & Brown, 2011). Schoenfeld describes how his research team's review of existing observation instruments influenced their development of the TRU Math observation protocol: "Ultimately, none of the [other] schemes jibed with our sense of what was central in good algebra teaching. . .Things we saw the teachers doing, that we judged to be important, were not reflected in the coding we did" (2013, p. 610). In other words, his research team deemed the validity of the existing instruments—built for other purposes—inadequate for capturing what they were interested in studying about good algebra teaching.

RELIABILITY OF THE OBSERVATION DATA

The focus and purpose of the observation system determines the methods of gathering, analyzing, and using data (Harvey, 2006; Vitiello & Hadden, 2014). As such, when designing or choosing an observation system it is important to consider the system's reliability for producing useful data. Because classroom observations can be quite unstructured, the reliability of the data can vary. For example, some observations are conducted to provide coaching and individual feedback on instruction and often take into account the uniqueness of a classroom's context; these fall into the category of unstructured classroom observations. While such unstructured observations can provide rich descriptions of practice and may be useful for developing theory or generating hypotheses about practice, they rarely allow comparisons across lessons or teachers and may result in "unfocused," "subjective," and/or "imprecise" observations that "may lead to premature judgments" (Harvey, 2006, p. 6; see also Pianta & Hamre, 2013). The unfocused nature of unstructured classroom observations, while perhaps beneficial for individual teachers, makes them difficult to apply consistently and reliably at scale, and thus less useful for understanding instructional trends (Schoenfeld, 2013).

Adding structure to unstructured observations can reduce individual bias and increase precision, reliability, and usefulness at scale. However, more structured observation systems also require advance design that may benefit from drawing on existing educational research (Harvey, 2006). For example, observers interested in

Table 1. Features of existing classroom observation tools

OBSERVATION PROTOCOL	PURPOSE	FOCUS	DATA COLLECTION	USE
Classroom Assessment Scoring System Available at: http://teachstone.com/the-class-system/	Designed to capture interactions linked to student academic, social, and self-regulatory development.	**CONTENT:** Not content specific **GRADE LEVEL:** PK–secondary **FOCUS DOMAINS:** » Classroom organization » Instructional support » Emotional support » Student engagement	**WHO OBSERVES:** Certified observers **WHAT DATA:** To assign codes, observers take detailed notes at the indicator level during classroom observation; review evidence gathered; and use detailed descriptions in the manual to code each dimension, scaled from 1–7. **WHEN:** For most projects and individual teacher feedback, 4 cycles per classroom within a year; for program-level decisions, 2–3 cycles per classroom.	Used for teacher feedback, goal-setting, and professional development.
Mathematical Quality of Instruction (MQI) Available at: http://isites.harvard.edu/icb/icb.do?keyword=mqi_training&tabgroupid=icb.tabgroup120173	Designed to provide scores for teachers on dimensions of classroom mathematics instruction.	**CONTENT:** Mathematics **GRADE LEVEL:** K–9 **FOCUS DOMAINS:** » Richness of mathematics » Working with students and mathematics » Errors and imprecision » Common Core-aligned student practices	**WHO OBSERVES:** Individuals complete online training to become MQI-certified observers. **WHAT DATA:** Observers view video, taking notes as needed, and assign ratings to short segments of videotaped lessons. Dimensions are scored differently, as follows: » 4-point rubric on 22 dimensions. » 2-point rubric on 1 dimension. » 5-point rubric on 10 whole-lesson dimensions. **WHEN:** Four video segments selected per teacher per year.	Used in research, teacher professional development, and evaluation.

continued on page 7

Table 1. Features of existing classroom observation tools *(continued)*

OBSERVATION PROTOCOL	PURPOSE	FOCUS	DATA COLLECTION	USE
CCSS Instructional Practice Guide Available at: http://achievethecore.org/content/upload/InstructionalPracticeGuide_MATH_K8_D_09192013.pdf	Designed to identify connections between Common Core-aligned lesson planning and classroom instruction.	**CONTENT:** Mathematics **GRADE LEVEL:** K–8 **FOCUS DOMAINS:** » Lesson reflects shifts required by CCSS-M. » Teacher employs instructional practices that allow all students to master the content of the lesson. » Teacher provides all students with opportunities to exhibit mathematical practices in connection with the content of the lesson.	**WHO OBSERVES:** Teachers, those who support teachers, others working to implement the CCSS-M **WHAT DATA:** Ratings based on whole lessons; 4-point rubrics on 16 dimensions **WHEN:** Some or most of the indicators and student behaviors should be observable in every lesson, though not all will be evident in all lessons; some actions may be viewed over the course of 2–3 class periods.	To facilitate reflective conversations between teachers and coaches about aligning content and instruction.
Teaching for Robust Understanding Available at: http://map.mathshell.org/materials/trumath.php	Designed to capture aspects of teaching believed to be consequential for students' development of robust algebraic understandings.	**CONTENT:** Mathematics **GRADE LEVEL:** K–12 **FOCUS DOMAINS:** » Mathematics » Cognitive demand » Access to mathematical content » Agency, authority, and identity » Uses of assessment » [Algebra-specific]	**WHO OBSERVES:** Observers vary by purpose, but may include trained researchers, professional development providers, and teachers. **WHAT DATA:** Episodes of up to 10 minutes are each coded separately on a 3-point rubric. Coders use sub-rubrics for different activity structures such as whole-class discussions, small group work, student presentations, and individual student work. **WHEN:** For research purposes, and to reflect teachers' practice generally, developers recommend ~6 scoring opportunities per teacher over a year, to look for consistency and growth patterns.	To date, primarily used for research purposes. Professional development tools (released in 2014) can be used by teachers reflecting on their own practice, coaches, or professional learning communities to enhance reflection and performance.

continued on page 8

Table 1. Features of existing classroom observation tools *(continued)*

OBSERVATION PROTOCOL	PURPOSE	FOCUS	DATA COLLECTION	USE
The 5x8 Card (SERP Institute) Available at: http://math.serpmedia.org/5x8card/	Designed to focus attention on student actions in order to support learning about shifts demanded by the CCSS Standards for Mathematical Practice.	**CONTENT:** CCSS Standards for Mathematical Practice **GRADE LEVEL:** K-8 **FOCUS DOMAINS:** Focus on 7 vital student actions observable in CCSS-M classrooms: 1. Equity requires participation. 2. Logic connects sentences. 3. Understanding each other's reasoning develops reasoning proficiency. 4. Revising explanations solidifies understanding. 5. Academic language promotes precise thinking. 6. ELLs develop language through explanations. 7. Productive struggle produces growth.	**WHO OBSERVES:** Principals; also relevant across actors with different responsibilities and expertise **WHAT DATA:** Principals observe and note evidence on card of vital student actions. **WHEN:** Not specified	SERP team members have used the 5x8 Card to organize professional development. A small deck of observations focused on teachers' instruction is currently under development.

studying teachers' questioning techniques may want to review the literature to understand how to identify and distinguish among higher-order and lower-order questioning strategies to make sure an observation tool can capture the desired distinctions and produce the desired data.

An additional design decision to consider about structured observation systems is the degree of inference required by observers. In some instances, observers may be asked to record factual information that does not require subjective interpretation—for example, checking a box when they observe a student presenting a mathematics solution to peers during whole-class discussion. Such checkbox/low-inference data may add consistency to the data collection, are easily quantifiable across classrooms, and may provide quick statistical information on instructional trends across teachers, schools, and a district. However, when classroom activities are reduced to "yes or no" checkmark indicators, many aspects of classroom learning are likely to be neglected, and more nuanced details on the variation or quality of the instruction are lost (Harvey, 2006).

The research indicates the importance of training (and re-training) observers to gather high-quality, consistent

data over time. Training is particularly important to help observers make consistent and reliable judgments about higher-inference categories of instruction; to do this, observers must have clear understandings of what constitutes the category of instruction they are observing and what specific classroom practices count as evidence. Accordingly, to achieve high levels of reliability when using a classroom observation protocol, many existing observation tools (e.g., MQI and CLASS) require users to practice to become certified observers familiar enough with the coding categories and criteria to be able to have a strong and consistent rationale for their coding. These observation systems draw on rating handbooks or manuals that clearly specify what counts as evidence for a particular rating, help structure the decision-making process, and "discourage consideration of irrelevant factors" (Milanowski et al., 2007, p. 5). Trained observers can gather the type of highly reliable, consistent data across time needed to inform ongoing implementation and system improvement.

Considerations for Math in Common Districts Implementing Classroom Observation Systems

The classroom observation design features highlighted in the previous section translate into five practical considerations for Math in Common districts to keep in mind as they select or develop and begin to implement their own systems for documenting teachers' instructional shifts in relation to the CCSS-M:

1. Choosing an observation tool
2. Choosing the observation sample
3. Scheduling the observations
4. Selecting and training the observers
5. Using the data

Below we describe each of these considerations for Math in Common districts.[2] Additionally, in Appendix A we provide a tool (modeled on Table 1) that may be useful for district representatives to complete as they review and reflect on their own observation tools.

CHOOSING AN OBSERVATION TOOL

As Math in Common districts are deciding whether to use/adapt an existing tool or develop a new one, it is useful to think through the three design principles outlined in the previous section: purpose of the classroom observation system, focus of the observations, and features of the observation system needed to produce reliable and useful data. As Schoenfeld (2013) and his research team discovered, existing tools may have both virtues and challenges for capturing exactly what one hopes to observe and measure. Because of this, homegrown or adapted versions of existing observation tools may provide data that is more useful for local policy decisions than existing tools. Schoenfeld (2013) offers three observations about using classroom observation systems to explore classroom behavior:

1. **Focus influences choice.** Depending on the intended focus of the tool, different aspects of instruction may play more or less central roles.
2. **Test in the real world.** No matter what instructional ideas are being studied, it is important to test theoretical ideas in the real world.
3. **Use variety.** Getting at the intended instructional focus (i.e., "what counts") requires multiple methods for gathering data and different perspectives.

Schoenfeld's reflections remind us that choosing any observation tool will depend on the district's intended focus, must be tested in the classroom, and will not necessarily provide sufficient information to answer all questions about instruction. Several of the articles included in the annotated bibliography provide useful guidance on the trade-offs required when using or adapting existing tools versus developing unique, context-specific tools (see, e.g., Education First, 2014; Joe, Tocci, Holtzman, & Williams, 2013; Milanowski et al., 2007; Stuhlman et al., 2014). Each of the Math in Common districts may want to carefully consider what data their observation tool will—and will not—provide to support district improvement efforts. Additionally, the New Teacher Project (2011) suggests five useful

[2] One additional consideration beyond the scope of this report is the groundwork needed to pave the way for the use of classroom observations in school districts. Because school district and teachers' union representatives may perceive the purpose and function of observations somewhat differently, it is important to address and discuss how and when observation systems are appropriate for use within a district setting.

Considerations for an effective observation system

"Instead of the crude checklists principals relied on in the past, leading school systems are equipping administrators and other evaluators with sophisticated observation instruments, often referred to as 'frameworks' or 'rubrics.' Some school systems are creating customized observation instruments while others are adopting or adapting commercially available ones. The new instruments enable observers to identify teaching practices along multiple dimensions and to classify practices along a continuum of performance levels, with the highest level of performance painting a picture of what excellent practice should look like." (Jerald, 2012a, p. 7)

questions to determine whether observation criteria and tools are likely to contribute accurate and useful results:

1. Do they cover the classroom performance areas most connected to student outcomes?
2. Do they set high performance expectations for teachers, or do they settle for minimally acceptable performance?
3. Are the performance expectations for teachers clear and precise?
4. Are they student-centered, requiring evaluators to look for direct evidence of student engagement and learning?
5. Are they concise enough for teachers and evaluators to understand thoroughly and use easily?

PRACTICAL IMPLEMENTATION TIP

As Math in Common districts are interested in understanding classroom instruction and documenting the implementation of the CCSS-M, it might be useful to examine observation systems from that perspective and choose or develop one that most closely aligns with this purpose and includes strong links to ideas included in the CCSS-M. While ease of use might be important, choosing a tool that only asks simple, check-box questions about whether students are engaged in doing CCSS-M or exhibiting mathematical practices may not suffice in providing the information needed to determine and describe variances in teachers' implementation of CCSS-M standards. Finding or developing a tool that allows for descriptive data about how teachers are engaging students and how students are exhibiting mathematical practices will likely provide more substantive insights into implementation successes and challenges. Choosing to use any particular observation tool—whether selected from an existing observation tool or developed within the district—will involve trade-offs and decisions about how the information will be used.

CHOOSING THE OBSERVATION SAMPLE

For most Math in Common districts, costs will prohibit systematic observations in all classrooms. The literature describes "differentiation" strategies in how districts are using classroom observations (White, 2014), and similarly Math in Common district personnel responsible for developing observation systems will need to identify an appropriate sample for conducting observations, based on overall goals. Although many researchers recommend conducting three to four observations per teacher to get an average sense of instruction in a particular teacher's classroom, they also recognize the associated complexity and costs of such intensive observations and recommend one to three observations per teacher per year as sufficient for supporting program-level decision-making (Jerald, 2012a; Shih, 2013; Hill & Grossman 2013; New Teacher Project, 2011; Vitiello & Hadden, 2014). Some researchers suggest that it may be more useful to differentiate observations (e.g., based on teacher experience) or select teachers known to

struggle more with particular instructional ideas or in particular school contexts to better understand variations in implementation and identify areas for needed professional development (Hill & Grossman, 2013; MET Project, 2013; White, 2014).

Pilot-testing in a few locations or with small groups of teachers will provide useful preliminary information on how well the observation system produces useful data about the sample group and whether the information might be broadly generalizable about instruction in the district in order to inform school and district decision-making. For example, Sartain et al. (2011) pilot-tested their observation system with a random sample of district elementary schools in order to ensure "that findings regarding implementation would be generalizable to other elementary schools across the city" (p. 6).

PRACTICAL IMPLEMENTATION TIP

As Math in Common districts will most likely be interested in understanding trends in implementation of the CCSS-M broadly, it may be wise to limit the number of observations for any particular teacher in favor of a broader sampling of instruction across classrooms or particular groups of teachers. The rationale for selecting the sample for classroom observations should be clearly specified and driven by central questions about instructional shifts, because the sample of observed teachers, grade levels, and schools will produce information about instruction specific to that sample. Sample selection must also be driven by capacity within the district and by the volume of data that will be produced; if the observation system produces too much data, it will be difficult for district staff to process and use the information for districtwide learning and improvement. As the classroom observations become more reliable and as the district improves its capacity to process and use classroom observation data to inform district decision-making, the sample for classroom observations can be increased.

SCHEDULING OBSERVATIONS

Observers in the Math in Common districts may want to identify particular observation windows during the school year (e.g., before and after a particular professional development training session), or randomly sample classrooms at different times during the year. Research suggests that notifying teachers of intended observations ahead of time does not change the distribution of observation ratings and that unannounced or random observations may ultimately reduce teacher anxiety, make scheduling easier for observers, and yield a more typical picture of instruction (Education First, 2014). However, random observations may also occur at inopportune times when instruction is influenced by particular events such as assessment preparation or elementary grade holiday festivities, thus yielding less useful information.

The window of time in which an observation occurs during a class period may also influence the data that is gathered, particularly in mathematics classes, as certain types of instruction may be present in some parts of lessons and not others. Some research suggests that short periods (e.g., 10–15 minutes) may be more productive for observations than full lessons, particularly as brief observations shift the observation paradigm from formal, "traditional, yet unproductive" observations to less formal, "frequent, shorter visits" that may be followed by some form of feedback to the observed teacher (Education First, 2014). There are pros and cons to observing different phases of lessons. For instance, an observer who stays throughout the entire class period can document how mathematical practices are used during whole-class instruction versus during small group activities, whereas an observer that stays for less time may not be able to compare and contrast different components of the lesson. Accordingly, it is important for the observation tool to capture the timing of the observation so that those attempting to make sense of collected observation data can understand findings in light of specific classroom contexts and patterns.

The iterative process of implementing an effective observation system

"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." (Joe et al., 2013, p. 2)

Several existing observation tools rely on videotaped observations of classroom instruction rather than real-time classroom observation. Videotaped observations enable researchers to pause and review instruction, which can increase the reliability of the data collection. However, videotaping may increase the overall time for data collection, since the observation requires both the live observation to videotape instruction and the subsequent coding and transfer of ratings to the observation tool. Practitioners will struggle with tighter time constraints and the desire for using data more quickly. Again, Math in Common districts will need to clearly specify their protocol for gathering classroom observation information and weigh the costs and benefits associated with decisions about how to implement the system.

PRACTICAL IMPLEMENTATION TIP

To observe classroom implementation of the CCSS-M, it may be useful for Math in Common district personnel to consult district and school calendars, and even testing windows, so the observations can be scheduled to optimize useful data collection. Additionally, the protocol for selecting when and for how long observations occur should be determined and documented ahead of time so observations are conducted similarly across teachers and observation data can be clearly understood and compared. Math in Common districts will need to balance the focus and grain size of their tool with the specifics of implementation. For example, they will need to make decisions that factor in the district's capacity for real-time versus videotaped classroom observations and observations of full math lessons versus 10–15 minute segments of instruction. They will also need to decide what sort of observation data will be the most useful to the district.

SELECTING AND TRAINING OBSERVERS

Math in Common districts will need to consider which staff they want to use to conduct the classroom observations. In the past, site principals or mathematics coaches were the individuals most frequently asked to conduct classroom observations (most often for purposes of teacher evaluation), but without adequate training they may not be ideal observers: some recent research has shown that the professional role of the person conducting the observation can influence the observation ratings. For instance, Sartain et al. (2011) reported on data from a Chicago Public Schools study showing that principals and trained classroom observers who were using the same tool (the Danielson Framework for Teaching) gave somewhat different ratings to various aspects of classroom instruction. Sartain and colleagues found that there were higher levels of agreement among raters on lower ends of the rating scale (i.e., unsatisfactory teaching), but that school principals were more likely to provide higher ratings than trained observers and to report that their evaluation ratings were influenced by the need to preserve relationships with teachers.

To document instructional shifts in CCSS-M, Math in Common districts may find it useful to employ a broad group of observers, including other teachers. Such broad involvement not only increases the observation capacity within a district system, but may also bolster teacher confidence in the data. This sort of broad involvement may also increase observers' professional learning

Training and Calibration for Classroom Observers

"The primary goals of observer training are to guide observers' understanding of the dimensions of the instrument and its rubrics and to give them the opportunity to hone their skill in applying the rubrics accurately. Without this step, the promise implicit in a shared definition…cannot be realized. To provide consistent and accurate observation scores, all observers must have the same understanding of what constitutes each level… [that] the system describes" (Joe et al., 2013, p. 8).

An effective training program for classroom observers should follow several important steps, as follows, led by a trained facilitator with extensive knowledge of the observation system:

1. Provide an overview of the process and help observers understand the purpose of conducting observations and the potential uses for the collected data.
2. Familiarize observers with the tool's overall areas of focus, rubric dimensions, and rating scales so that they can become familiar with the vision for effective instruction that underlies the tool, and the differences and relationships between the central ideas. Before attempting to use the tool, observers will need to have a thorough grasp of the observation tool's rationale, organization, and language. Allow observers to ask and discuss clarifying questions about how each of a rubric's components is described.
3. Help observers recognize their own biases and how such biases might influence interpretations of rubric dimensions or introduce judgment errors.
4. Guide observers to understand the range of performance the instrument describes for the elements of practice under each dimension. Describe how observers would recognize levels of performance.
5. Train observers on how to use the tool. For example, describe how they would take notes to support subsequent coding, record evidence of performance, or use a rating tool during real-time observation.
6. Have observers practice observing, collecting evidence, and connecting evidence to the tool so they begin to calibrate their own observations against the observation tool. It may be useful to have observers start practicing with video segments of instruction, so that they can pause and review particular segments of instruction in a low-stakes environment. Selected videos should reflect authentic teaching practice and a range of practices, levels of teaching quality, grades, and demographics so that observers become adept at applying the ratings across diverse classroom contexts.
7. Practice interpreting evidence relative to the rating scale by sharing ratings with other observers; discuss rationales gathered during observations (e.g., objective evidence and why a behavior did or did not merit a particular score) and rating conventions (e.g., What do you do if you do not have any evidence for a performance dimension? How do you interpret words like "consistently" or "frequently" when used in the rating scale?).
8. Continue to practice rating samples of performance. While starting practice with video examples may be useful for beginning observers, "live practice" with the coding tool may better support observers to transfer observation and coding skills to live classroom settings.
9. Conduct a certification assessment (i.e., where observers are required to match the ratings of an "expert" group in order to pass) and release observers gradually to conduct independent scoring after they have been certified. Additionally, require periodic re-certification to confirm that skills are maintained over time. Periodic re-certification might take the form of "deep-dive" training where a group of observers focus on a specific dimension, one-on-one coaching, paired observations of live or video-recorded lessons, or group calibration sessions.

Source: Adapted from Jerald, 2012a; Joe et al., 2013; McClellan, 2013; Milanowski, et al., 2007.

Aligning observation systems to the CCSS

"States and districts have learned a great deal in the last few years about how to create better teacher development and evaluation systems. But there's still much to learn as these systems are implemented and improved over time and aligned to new expectations for students. One of the most exciting prospects is aligning teacher development and evaluation systems to the Common Core State Standards. As they move forward, states and districts should commit to measurement but hold lightly to the specific measures as the field continues to gain new knowledge." (MET Project, 2013, p. 8)

about instruction and instructional shifts by enhancing their knowledge, skills, and ability to notice certain instructional features of interest and by supporting conversations about classroom activities and instruction (Education First, 2014; Harvey, 2006; Hill & Grossman, 2013; Sherin, Jacobs, & Philipp, 2011).

Math in Common districts should also be aware of the complexity (as documented in the research literature) of conducting classroom observations that are consistent, reliable, unbiased, and non-subjective. Accordingly, training the staff that will conduct the classroom observations is critical to the success of the observation protocol:

> Without training of the observer…many problems may occur. Anyone who undertook a classroom observation would recognize that there is a definite risk for both observation and interpretation processes to be partly subjective and biased. With training and experience, it is possible to conduct observation with the right state of mind, with a more passive and neutral engagement, getting the whole picture with no prejudice as well as focusing on specific events to collect evidence. (Harvey, 2006, p. 5)

Training to conduct observations should aim to erase biases and support reliable, consistent ratings regardless of who is gathering the observation data. Ultimately, such training can enable observers "to appreciate the benefits and limitations of the observation process [and] to recognize what is realistic and practical to do during observation" (Harvey, 2006, p. 8).

Additionally, like any form of professional learning, training to conduct classroom observations should not be considered a one-time activity. Observers need the knowledge, skills, and tools to do the job well (Jerald, 2012a) and to avoid changing their ratings across time. Training will need to be ongoing because ratings may fluctuate (i.e., "drift") over time as observers gain knowledge and feed that knowledge back into the system to support improvements to the observation system. To prevent rating drift it is important to continue calibration and training efforts (see the sidebar on Training and Calibration for Classroom Observers for guidance) over time to maintain reliability (Jerald 2012a; Vitiello & Hadden 2014; Bain, 2010; Sartain et al., 2011).

PRACTICAL IMPLEMENTATION TIP

Most Math in Common districts will not be able to devote full-time staff to conducting all observations. However, through training and shared experiences, districts can develop a group of "technically competent" observers who can use observation tools reliably and objectively to observe instructional shifts related to the CCSS-M and systematically document changes and improvements. Structured training and calibration processes should be well documented and built into the classroom observation system.

USING THE DATA

For Math in Common districts, the primary motivation for implementing a plan for observing mathematics classroom instruction is to improve instruction that is consistent with the Common Core State Standards. Teachers will need specific pedagogical support around

the standards for mathematical practice in order to adjust instruction accordingly and implement the CCSS-M successfully. While implementing a classroom observation system involves significant challenges, time, and costs, a high-quality, rigorous observation system can produce valuable formative data that documents teachers' classroom instruction and provides information to support teachers' further learning and implementation of the CCSS-M.

Additionally, by analyzing observation data across groups of teachers, districts can get a more general story about CCSS-M implementation outcomes and instructional shifts. Valid and reliable data also enables district stakeholders to examine whether district investments are effectively supporting teacher instruction and student achievement. (Less rigorous observation systems can also produce useful information, although users should be more skeptical about drawing general conclusions from such data.)

So how should the Math in Common districts approach the idea of using observation systems? The litmus test for the observation system should typically be whether, and in what ways, the information that is collected through the observations is *useful* to school leaders, district leaders, and teachers. If the information that is being gathered is valid and reliable, Math in Common teams will have increasing confidence that the instructional patterns they are observing in smaller samples are strongly associated with the instructional patterns for the district as a whole. If those patterns are indicating consistently strong instruction, district teams will have the leverage to further support that same type of strong instruction in more sites and classrooms; if those patterns indicate instruction that is consistently less than optimal, district teams will have an opportunity to reflect on why the instruction is not working well, and reposition support accordingly. Conversations about using observation data to inform instructional support can happen in professional learning communities, in coaching programs, and in planning for subsequent teacher professional development offerings. Regardless of where these conversations occur, districts can use valid and reliable observation data to produce rich, descriptive information about instruction that helps school systems arrive at shared understandings of successful implementation of the CCSS-M and strategies for improvement.

PRACTICAL IMPLEMENTATION TIP

Given that implementing an observation system is costly and takes substantial staff time, it is essential that the Math in Common teams use the information from the observational work strategically. The development of recursive systems where information is used thoughtfully might best be done one grade at a time across a few sites, and scaled up over several semesters. Data use, of course, is not a one-size-fits-all undertaking; findings from and uses of classroom observation data in each of the Math in Common districts will be context- and tool-dependent. How one district operationalizes their use of classroom observation data to support subsequent investments across time may differ significantly from others, depending on the nature of the district's questions about instruction, the design of the observation system, and the recursive uses within their school system.

Annotated Bibliography on Classroom Observations

REFERENCES

Education First (2014). *Classroom observations: Measuring teachers' professional practices.* Washington, DC: Author. Available from: http://www.education-first.com/files/EdFirst_Eval-ClassroomObservations-Nov11.pdf

Abstract: Culling from research and the experience of leading districts and states, the report aims to answer two questions: (1) What do high-quality classroom observations look like? (2) What must states and districts do better if they want to improve teachers' skills? Four recommendations are offered for using classroom observations to measure teachers' professional practice: (1) ensure systemwide understanding of effective practice; (2) build credibility for the observations by certifying observers and providing teachers with high-quality feedback; (3) make observations useful and manageable for teachers and principals; and (4) monitor observation results for quality and build a system of continuous improvement. The report concludes that high-quality classroom observations take a lot of work, but—if executed correctly—can result in substantial pay-off, particularly when states and districts also commit the full suite of supports needed to implement high-quality observations, including training, time, resources, and new policies.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis.* Washington, DC: National Comprehensive Center for Teacher Quality. Available from: http://files.eric.ed.gov/fulltext/ED521228.pdf

Abstract: This research synthesis examines how teacher effectiveness is currently measured. By evaluating the research on teacher effectiveness and the different instruments used to measure it, including classroom observations (discussed on pages 20–25), this research synthesis contributes to the discussion of appropriate rigor and relevance of measures for different purposes (e.g., formative vs. summative evaluation). The findings are presented along with related policy implications. In addition, the synthesis describes how various measures have been evaluated, explains why certain measures are most suitable for certain purposes (e.g., high-stakes evaluation versus formative evaluation), and suggests how the results of the study might be used to inform the national conversation about teacher effectiveness. A comprehensive definition of the components and indicators that characterize effective teachers is provided, extending this definition beyond teachers' contribution to student achievement gains to include how teachers impact classrooms, schools, and their colleagues as well as how they contribute to other important outcomes for students. Through this synthesis, the National Comprehensive Center for Teacher Quality hopes to provide some practical guidance on how best to evaluate teacher effectiveness.

Harvey, P. (2006). *Improving teaching observation practice in the learning and skills sector: A literature review.* Norwich, England: City College Norwich. Available from: http://theresearchcentre.co.uk/files/docs/publications/rs3916.pdf

Abstract: Since 2000, classroom observation has been an increasing part of the United Kingdom's teacher appraisal and inspection systems, used within three different contexts: for research purposes, as a part of assessment to monitor the quality of teaching within an organization, and for teacher professional development. This study focused primarily on the last category of classroom observation and feedback, intended to support novice teachers in the developmental phase of their new profession. The study reviewed the literature on best practice of classroom observation and feedback procedures and identified the strengths and any [challenging] issues of the present practice. The literature review covers issues related to and best practices

regarding observation methodology, procedures, and post-lesson analysis and feedback. The project also examined other critical procedures in classroom observation including observation of experienced teachers, videoing, and paired observations.

Hill, H. C., & Grossman, P. (2013, Summer). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review*. Cambridge, MA: Harvard Publishing Group. Available from: http://cepr.harvard.edu/cepr-resources/files/news-events/ncte-hill-grossman-learning-from-teacher-observations.pdf

Abstract: In this article, Heather Hill and Pam Grossman discuss the current focus on using teacher observation instruments as part of new teacher evaluation systems being considered and implemented by states and districts. The article argues that if these teacher observation instruments are to achieve the goal of supporting teachers in improving instructional practice, they must be subject-specific, involve content experts in the process of observation, and provide information that is both accurate and useful for teachers. The authors discuss the instruments themselves, raters and system design, and timing of and feedback from the observations. The paper concludes by outlining the challenges that policymakers face in designing observation systems that will work to improve instructional practice at scale.

Jerald, C. (2012a). *Ensuring accurate feedback from observations: Perspectives on practice*. Seattle, WA: Bill and Melinda Gates Foundation. Available from: https://docs.gatesfoundation.org/Documents/ensuring-accuracy-wp.pdf

Abstract: This brief offers examples and lessons from leading states, districts, charter management organizations, and other education organizations working to provide teachers with accurate feedback from observations. The report identifies three broad areas for action to ensure accurate observations: (1) build observers' capacity to conduct accurate observations, (2) create conducive conditions for observing accurately in the field, and (3) monitor observations periodically to ensure quality. The report includes "threat assessment" questions for assessing potential problems and building capacity to respond to the three action areas. The report concludes that there is not yet evidence to suggest a single way of providing accurate feedback from observations, and that school systems should design coherent solutions that fit their particular needs and select strategies based on a logical theory of action about how those strategies can ensure accurate observations.

Jerald, C. (2012b). *Movin' it and improvin' it! Using both education strategies to increase teaching effectiveness*. Washington, DC: Center for American Progress. Available from: http://cdn.americanprogress.org/wp-content/uploads/issues/2012/01/pdf/movin_it_improvin_it.pdf

Abstract: The report highlights two very different types of strategies for boosting teaching effectiveness and shaping teacher evaluation reform. The first strategy evaluates a teacher's effectiveness at a point in time then seeks to attract and keep teachers with higher effectiveness while removing teachers with lower effectiveness and raise the average level of effectiveness over time. The second strategy seeks to provide all teachers with useful feedback following classroom observations or use the results of evaluation to individualize professional development for teachers to increase the average effectiveness in the workforce. The report reviews recent research on teacher professional development and concludes that the two strategies, when used together, are likely to maximize increases in teaching effectiveness. The author advises districts to conduct comprehensive audits of all of their investments in professional development to determine whether the investments provide real opportunities for teachers to improve and to take every step possible to ensure that feedback teachers receive from evaluations is as valuable as teachers have been promised.

Joe, C. N., Tocci, C. M., Holtzman, S. L., & Williams, J. C. (2013). *Foundations of observation: Considerations for developing a classroom observation system that helps districts achieve consistent and accurate scores.*

Princeton, NJ: Education Testing Service and Seattle, WA: The Measures of Effective Teaching Project, Bill and Melinda Gates Foundation. Available at: http://www.metproject.org/downloads/MET-ETS_Foundations_of_Observation.pdf

Abstract: This report provides information on processes states and school districts can use to help ensure high-quality data collection during teacher observations. The paper draws on the knowledge and expertise of the Education Testing Service, gained from (1) designing and implementing scoring processes for the Measures of Effective Teaching project, and (2) from experience with other observation systems. The report outlines 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. The report concludes by discussing 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, especially as stakes assigned to observation data increase and have the potential to become more litigation-prone.

Leo, S. F., & Coggshall, J. G. (2013). *Creating coherence: Common Core State Standards, teacher evaluation, and professional learning*. Washington DC: Center on Great Teachers and Leaders, American Institutes for Research. Available from: http://www.gtlcenter.org/sites/default/files/CreatingCoherence.pdf

Abstract: This brief from the Center on Great Teachers and Leaders introduces an approach to creating coherence among three instructional reforms: implementation of the Common Core State Standards, new standards-based teacher evaluation systems, and job-embedded professional learning. This approach is based on a set of Core Instructional Practices that describe the kinds of instruction teachers need to be implementing in their classrooms to help their students meet the Common Core State Standards. The report provides a broad overview of the reforms and their intersections, then lays out an approach to improving coherence. Subsequently, the authors illustrate the approach with three teacher evaluation frameworks: Charlotte Danielson's Framework for Teaching Evaluation Instrument, Robert Marzano's Teacher Evaluation Model, and Robert Pianta's Classroom Assessment Scoring System Rubric. The report demonstrates areas of alignment between the Core Instructional Practices and any research-based professional practice framework and also identifies opportunities to create better coherence between the frameworks and Core Instructional Practices. The report concludes with a description of implications of this work for professional learning, particularly discussing thinking and learning about practice as a way to create coherence with an actionable plan for implementation.

Marzano, R. J., & Toth, M. D. (2014). *Teaching for rigor: A call for a critical instructional shift*. West Palm Beach, FL: Learning Sciences International. Available from: http://www.broward.k12.fl.us/talentdevelopment/news/teacher_links/Teaching-for-Rigor-20140318.pdf

Abstract: This paper offers a model of essential classroom strategies to support the instructional shifts in pedagogy needed in an environment of academic rigor for all students. The paper describes data from a large sample of classroom observations and analysis by Learning Sciences Marzano Center that document the pedagogical strategies teachers are currently using in their classrooms. The paper finds that the majority of teachers are not adequately prepared to make the critical instructional shifts necessary to meet the requirements for rigor in college and career readiness standards. Finally, the paper describes a model of instruction, focusing on 13 essential classroom strategies for achieving rigor, to refine and supplement teacher instructional skills to meet rigorous new standards.

The Measures of Effective Teaching Project (2013). *Feedback for better teaching: Nine principles for using measures of effective teaching*. Seattle, WA: Bill and Melinda Gates Foundation. Available from: http://www.metproject.org/downloads/MET_Feedback%20for%20Better%20Teaching_Principles%20Paper.pdf

Abstract: To help states and districts navigate the work of implementing feedback and evaluation systems that support teachers, this brief highlights a set of nine guiding principles from the Bill and Melinda Gates Foundation to inform the design and implementation of high-quality teacher support and evaluation systems. The paper is based on three years of work by the Measures of Effective Teaching project, its partners, and other leading school systems and organizations. The report describes three overarching imperatives for implementing high-quality teacher support and evaluation systems, including: measure effective teaching; ensure high-quality data; and invest in improvement. These three imperatives are linked in a cyclical fashion to demonstrate the need for well-designed evaluation systems to continue to improve over time.

Milanowski, A. D, Prince, C. D, & Koppich, J. E., (2007). *Observations of teachers' classroom performance: Guide to implementation, resources for applied practice*. Washington, DC: U.S. Department of Education, Office of Elementary and Secondary Education, Center for Educator Compensation Reform. Available from: http://cecr.ed.gov/pdfs/guide/CECRTeacherObservationModel.pdf

Abstract: This report provides guidance for constructing an observation system to ensure it measures the right things, produces valid and reliable measurements, provides tools to help educators improve performance and is accepted by those whose performance is being measured and by those doing the measuring. Additionally, detailed recommendations are provided to support districts in several aspects of using classroom observations, including

- developing rating scales,
- defining evidence and how it will be collected,
- creating an analytic assessment process,
- using multiple evaluators,
- training evaluators for consistency, and
- monitoring evaluators' performance and holding evaluators accountable.

The report includes examples of observation instruments from several districts. Although the report was written to discuss classroom observations for the purpose of informing educator compensation decisions, many of the highlighted findings and recommendations are highly relevant to classroom observations more generally.

New Teacher Project. (2011). *Rating a teacher observation tool: Five ways to ensure classroom observations are focused and rigorous*. Brooklyn, NY: Author. Available from: http://files.eric.ed.gov/fulltext/ED544422.pdf

Abstract: The report begins with background on six proposed design standards that any effective teacher evaluation system should meet, and describes the role of including both objective student learning data and subjective observation data of teachers' classrooms (focused on skills that can be directly observed). A second part of the report focuses on a series of questions to determine whether observation criteria and tools are likely to contribute to accurate evaluation results. A third part of the report includes a tool to assess the quality of any observation rubric.

Pianta, R., & Hamre, B. (2013). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher 38*(2), 109–119. Available from: http://147.226.7.60/-/media/WWW/DepartmentalContent/Teachers/PDFs/109full.pdf

Abstract: The authors advance an argument that placing observation of actual teaching as a central feature of accountability frameworks, teacher preparation, and basic science could result in substantial improvements in instruction and related social processes and a science of the production of teaching and teachers. Teachers' behavioral interactions with students can be (1) assessed observationally using standardized protocols, (2) analyzed systematically with regard to sources of error, (3) validated for predicting student learning, and (4) changed (improved) as a function of specific and aligned supports provided to teachers; exposure to such

supports is predictive of greater student learning gains. These methods have considerable promise; along with measurement challenges—some of which pertain to psychometrics, efficiency, and costs—they merit attention, rigorous study, and substantial research investments.

Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011). *Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation* (Research Report). Chicago, IL: Consortium on Chicago School Research. Available from: https://ccsr.uchicago.edu/sites/default/files/publications/Teacher%20Eval%20Report%20FINAL.pdf

Abstract: The report summarizes findings from a two-year study of Chicago's Excellence in Teaching Pilot, designed to drive instructional improvement by providing teachers with evidence-based feedback on their strengths and weaknesses. The pilot consisted of training and support for principals and teachers, principal observations of teaching practice conducted twice a year using the Charlotte Danielson Framework for Teaching, and conferences between the principal and the teacher to discuss evaluation results and teaching practice. The authors found that the pilot was an improvement on the old evaluation system and worked as it was designed and intended, introducing an evidence-based observation approach to evaluating teachers and creating a shared definition of effective teaching. However, the new system also faced a number of challenges, including weak instructional coaching skills and lack of buy-in among some principals. The final chapter of the study provides a design guide for districts and unions attempting to revitalize teacher evaluation systems. The authors conclude that building a successful evidence-based teacher evaluation system requires an intentional, long-term commitment.

Schoenfeld, A. H. (2013). Classroom observations in theory and practice. *ZDM, the International Journal of Mathematics Education, 45*, 607–621. Available from: http://ncm.gu.se/media/ncm/dokument/classroom_obs.pdf

Abstract: This essay provides an overview of the process that led to the development of the TRU Math rubric. The author begins by discussing the complexities of constructing a classroom analysis scheme for empirical use even when a general theory regarding teacher decision-making is available. Next, the author presents the scheme that was developed (the TRU Math rubric) and the necessary and sufficient set of dimensions for the analysis of effective classroom instruction. Schoenfeld describes the process and results of looking at a wide range of schemes that other researchers or professional developers had constructed for the analysis of classroom interactions. The paper concludes with three observations on using classroom observation protocols.

Shih, J. C. (2013). How many classroom observations are sufficient? Empirical findings in the context of a longitudinal study. *Middle Grades Research Journal, 8*(2), 41–49. Available upon request or via Google search.

Abstract: One method to investigate classroom quality is for a person to observe what is happening in the classroom. However, this method raises practical and technical concerns such as how many observations to collect, when to collect these observations and who should collect these observations. The purpose of this study is to provide empirical evidence to address these concerns using a particular middle school mathematics classroom observation tool. Findings suggest that raters trained to use this particular measure required three observations to consistently capture habitual classroom environments. Implications for investigating classroom quality using this and other classroom observation tools should be guided by decisions about the specific purpose of the observation tool, as well as budget and practical considerations.

Stuhlman, M., Hamre, B., Downer, J., & Pianta, R. (2014). *A practitioner's guide to conducting classroom observations: What the research tells us about choosing and using observational systems.* Charlottesville, VA: University of Virginia, Curry School of Education. Available

from: http://curry.virginia.edu/resource-library/practitioners-guide-to-classroom-observations

Abstract: Educational leaders, who prepare, evaluate, and support teachers have many responsibilities, but none more important than supporting teachers in delivering high-quality instruction to students in their classrooms. To facilitate high-quality teaching practices most effectively and efficiently, teacher preparation programs, principals, schools systems, and all those who work with and mentor teachers need tools that facilitate progress towards this goal. In this report, the authors document the ways in which one such tool, standardized observational assessments, can guide educational organizations, promoting effective teaching practices that enhance students' social and academic development. The report is designed to provide school personnel with research-based information about using observational methodology in five key areas:

- Why Should We Use Classroom Observation
- What Should Classroom Observation Measure
- How to Select the Right Classroom Observation Tool
- How to Use Classroom Observation Most Effectively
- How Classroom Observations Can Support Systematic Improvement in Teacher Effectiveness

Vitiello, V. E. & Hadden, D. S. (2014). *CLASS system implementation guide: Aligned improvement solutions.* Charlottesville, VA: Teachstone Training, LLC. Available from: http://teachstone.com/the-class-system/

The interactions children have with teachers and peers are the single most important classroom influence on learning. The Classroom Assessment Scoring System (CLASS) observation measures were developed to accurately capture the interactions most closely linked to academic, social, and self-regulatory development in young children and students from birth through secondary school. Because the CLASS measures define effective interactions in specific, behavioral terms, they also serve as an actionable foundation for teacher professional development. The purpose of this CLASS System Implementation Guide is to help states, counties, districts, and programs understand how to use the CLASS system to observe and improve teacher-child interactions. Guidelines for measuring effective teacher-child interactions and improving teaching and learning are provided.

White, T. (2014). *Evaluating teachers more strategically: Using performance results to streamline evaluation systems.* Palo Alto, CA: Carnegie Foundation for the Advancement of Teaching. Available from: http://cdn.carnegiefoundation.org/wp-content/uploads/2014/12/BRIEF_evaluating_teachers_strategically_Jan2014.pdf

The issue brief explores differentiation strategies in nine districts, two charter management organizations, and three states (Tennessee, Delaware, and Ohio), reporting that many of these school systems have embraced differentiation strategies as a way to conserve teacher evaluation resources or to deploy existing resources more efficiently. The report describes several formats for observation (e.g., walkthroughs or "partials" compared to more formal full-length classroom observations), and describes how organizations are rethinking frequency of observations and the mixture of formal and informal observations. A table in the report compares several features of observation systems across the school systems, such as who conducts the observations. A report endnote provides a link for online cost calculator to help district employees and members of the K–12 community understand the different components of designing a district's teacher evaluation system.

ADDITIONAL REFERENCES CITED IN REPORT

Bain, A. (2010). A longitudinal study of the practice fidelity of a site-based school reform. *The Australian Educational Researcher, 37*(1), 107–123.

Cohen, D. K. (1990). A revolution in one classroom: The case of Mrs. Oublier. *Educational Evaluation and Policy Analysis, 12*(3), 327–345.

Cohen, D. K., & Hill, H. C. (2001). *Learning policy: When state education reform works*. New Haven: Yale University Press.

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education, 119*(3), 445–470.

McClellan, C. (2013). *What it looks like: Master coding videos for observer training and assessment*. Seattle, WA: Bill and Melinda Gates Foundation.

McDonald Connor, C. (2013). Commentary on two classroom observation systems: Moving toward a shared understanding of effective teaching. *School Psychology Quarterly, 28*(4), 342–346.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F., (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417–458.

Rosenshine, B. (1970). Interpretive study of teaching behaviors related to student achievement (Final Report). Washington, DC: National Center for Educational Research and Development (DHEW/CE).

Rosenshine, B., & Furst, N. (1973). The use of direct observation to study teaching. In R. M. W. Travers (Ed.). *Second handbook of research on teaching* (pp. 122–183). Chicago: Rand McNally.

Sherin, M., Jacobs, V., & Philipp, R. (2011). *Mathematics teacher noticing: Seeing through teachers' eyes*. New York: Routledge.

Spillane, J. P. (2000). Cognition and policy implementation: District policymakers and the reform of mathematics education. *Cognition and Instruction, 18*(2), 141–179.

Stewart, B. E. (2006). *Value-added modeling: The challenge of measuring educational outcomes*. New York, NY: Carnegie Corporation

University of Michigan. (2006). *A coding rubric for measuring the mathematical quality of instruction* (Learning Mathematics for Teaching Technical Report LMT1.06). Ann Arbor, MI: University of Michigan, School of Education.

Youngs, P. (2013). *Using teacher evaluation reform and professional development to support Common Core assessments*. Washington, DC: Center for American Progress.

Appendix A.
Methods for Annotated Bibliography

KEYWORDS AND SEARCH STRINGS USED IN THE SEARCH

"Classroom observation" AND "protocols" OR "implementation" OR "purpose" OR "policy implementation" OR "teacher effectiveness."

SEARCH OF DATABASES

EBSCO Host, Google, and Google Scholar.

CRITERIA FOR INCLUSION

When reviewing resources, we considered three main factors:

- **Date of the publication:** The most current information is included, except in the case of nationally known seminal resources.
- **Source and funder of the report/study/brief/article:** Priority is given to IES, nationally funded, and certain other vetted sources known for strict attention to research protocols.
- **Methodology:** Sources include randomized controlled trial studies, surveys, self-assessments, literature reviews, and policy briefs. Priority for inclusion generally is given to randomized controlled trial study findings, but the reader should note at least the following factors when basing decisions on these resources: numbers of participants (Just a few? Thousands?); selection (Did the participants volunteer for the study or were they chosen?); representation (Were findings generalized from a homogeneous or a diverse pool of participants? Was the study sample representative of the population as a whole?).

Appendix B.
Classroom Observation Analysis Tool

OBSERVATION PROTOCOL	PURPOSE	FOCUS	DATA COLLECTION	USE
[Insert information on your own district observation protocol or your adapted version of an existing protocol]		**CONTENT:** **GRADE LEVEL:** **FOCUS DOMAINS:**	**WHO OBSERVES:** **WHAT DATA:** **WHEN:**	