Center for American Progress

THE ASSOCIATED PRESS/JOHN RUSSELL

Implementing Observation Protocols

Lessons for K-12 Education from the Field of Early Childhood

Robert C. Pianta May 2012

WWW.AMERICANPROGRESS.ORG

Center for American Progress

Implementing Observation Protocols

Lessons for K-12 Education from the Field of Early Childhood

Robert C. Pianta May 2012

Contents

Introduction and summary

While it might seem counterintuitive, at least some of the answers to turning around our nation's struggling K-12 public schools can be found at the nearest preschool.

At a time of considerable urgency and demand for improvements in our nation's schools, particularly when it comes to evaluating the effectiveness of teachers, there is no need to reinvent the wheel. Instead of looking to the development and implementation of new educational models and methodologies, K-12 educators would do well to learn from the lessons and experience accrued by their counterparts in the early childhood sector, specifically when it comes to teacher performance evaluation.

There is no shortage of debate on the challenges and promises of teacher performance evaluation as the reauthorization of the Elementary and Secondary Education Act of 2001, also known as No Child Left Behind, proceeds and as states seek to implement reforms. Unfortunately, there is precious little precedent for the use of performance evaluation of teachers in the K-12 education setting, at least good performance evaluation.[1] The well-documented shortcomings of existing evaluation methods from principal "drive-by" observations to hiring interviews to tenure reviews and more all lead to the same conclusion—nearly every teacher "passes" whatever "test" they face. The problem is that the "tests" themselves do not discriminate good performers from poor performers and make virtually no connection between these "tests" and student achievement, professional development, or incentives to improve.

Relying on the status quo for teacher performance evaluation wastes time and energy—performance metrics are nonexistent or not valid and there is little to no linkage among the key components of most evaluation and performance-improvement systems. As practiced now teacher evaluation is a nonsystem with a lot of moving parts of dubious value and very little connection among them.

Some measure of teachers' classroom practices, usually in the form of observation, is at the core of nearly every proposal and early-stage rollout of the next generation of teacher performance evaluation efforts in districts and states.[2] Typically coupled with estimates of teachers' contributions to student gains on achievement tests as well as with other indicators of performance, observation of teachers' classroom practices is a cornerstone of this new wave of assessment. To ensure that an evaluation system is capable of providing teachers with the actionable feedback needed to improve, solid information is paramount. Clearly, high-quality classroom behavior and practices are at the core of any definition of "effective teaching" and what most teachers would identify as the manner in which they contribute to student learning.

It is sensible to think that observational assessment of teachers' classroom behavior would be a central component of any evaluation system since teachers' behaviors and interactions are students' most direct experience of teaching. Yet like most initiatives in education reform, observation is subject to implementation and policy challenges that could very well hinder its ultimate benefits. The short list of challenges include: technical issues in defining and measuring teaching behavior; gathering information about a teacher through consistent and reliable observation; ensuring that the behaviors observed really matter for student learning (for example, validity of the observation); determining how observations connect to high-stakes consequences such as tenure and professional development; and a host of support and infrastructure requirements needed to roll out sound observation efforts on a large scale.[3] Yet there are too few models of how to do observation well in the K-12 sector. But there is one sector where we have more than two decades of widespread application of classroom observation from which to draw lessons: early childhood education, which is the focus of this paper.[4]

Like most initiatives in education reform, observation is subject to implementation and policy challenges that could very well hinder its ultimate benefits.

This report draws from decades of experience using observation in early childhood education, which has implications for administrative decisions, evaluation practices, and policymaking in K-12. Early childhood education has long embraced the value of observing classrooms and teacher-child interactions. In early childhood education the features of the settings in which children are served are the hallmarks of quality. These features can include health and safety considerations, the materials and physical layout of the space, and the interactions that take place between adults and children—such as conversations, emotional tone, or physical proximity. Standardized observations of these early childhood education features in turn yield metrics that are used in state and federal policy, program-improvement investments, and the credentialing of professionals[5]—all uses that K-12 education is now considering.

This paper examines lessons learned from observation in early childhood education that may be helpful as states and districts begin implementing more rigorous observation protocols for K-12 teachers. Although these lessons apply to all grades, they may be particularly relevant for K-3 as assessment of student performance using standardized achievement tests is most challenging in those grades. These lessons focus on the importance of standardization, trained observers, methods for ensuring the validity and reliability of the instruments, and the use of observational measures as a lever to produce effective teaching. These lessons form the basis for the following recommendations:

- Any measure must provide information in the form of metrics that clearly differentiate those being assessed. Observation is no exception—thus observation is a form of measurement and assessment consisting of codes and benchmarks that must be applied rigorously, just as they are in assessments of student performance.

- Observations used in systems of decision making and performance improvement must adhere to standardized procedures. There are three components of standardization that are key elements for evaluating any observation instrument and its implementation—training protocol, parameters around observation, and scoring directions.

- The technical properties of observational protocols and scoring systems are fundamental for their use. Reliability is one of these properties and pertains to the level of error or bias in the scores obtained. It is critical that users select tools that have documented reliability for use across observers, teachers, time, and situations. Effective training programs for observers help to ensure raters are consistent with one another as they make ratings. Similarly, including periodic "drift" testing at predetermined intervals will help to improve the degree to which raters remain consistent with scoring protocols and with each other.

- Any observation of teacher performance must show empirical relations with student learning and development if the use of observation is expected to drive improvement in student outcomes. Selecting an observation system that includes validity information cannot be overstated.

- Pragmatically, observation takes time and different systems of observation require different time commitments. The amount of observer time available can be an important practical consideration when selecting an observational system.

In general the more ratings a school or district is able to obtain and aggregate, the more stable an estimate of typical teacher practices will result.

- Observations can identify teacher classroom behaviors that matter for students, can describe typical teacher practices, can show how a given classroom or teacher compares with a national or district average, can forecast the likely contribution of a teacher to children's learning, or can document improvement in teachers' practices in response to professional development. Users, however, must be cautious to not overstep the appropriate use of observational instruments in their enthusiasm to apply them in any and all circumstances.

- Observations can be used in both accountability and program-improvement applications. Importantly, policy and program investments over time can change the typical distribution of scores as teachers, classrooms, and programs improve, and as a consequence it can be necessary to periodically "raise the bar" on performance standards or cutoff scores.

- Feedback to teachers is most effective when it is individualized and highly specific, focused on increasing teachers' own observation skills, promotes self-evaluation, and helps teachers see and understand the impact of their behaviors more clearly.

Note: To better make our point, we've employed the technique of using fictional situations throughout this paper to illustrate specific points that further our overall argument that the use of early childhood education observational evaluation methods have value for K-12 education.

Large-scale use of standardized observation protocols for early childhood settings and teachers

This section describes large-scale work being done in the observation of teachers and classroom settings in early childhood education. Most of the discussion focuses on two prominent observation systems—the Early Childhood Environment Rating Scale, or ECERS,[6] and the Classroom Assessment Scoring System, or CLASS.[7] We present explicit descriptions of observation use in the monitoring, accountability, and professional development framework of Head Start, in statewide programs for children from birth through five years of age, and in various states' Quality Rating and Improvement Systems[8] (analogous to Human Capital Management Systems in K-12). In addition, we describe uses related to high-stakes accountability decisions, program improvement, and identifying specific challenges and solutions.

ECERS: Early Childhood Environment Rating Scale

The suite of Environmental Rating Scales, or ERS, developed in the late 1970s and 1980s by researchers Richard Clifford, Thelma Harms, and colleagues have been nothing short of foundational to the development of the early childhood education infrastructure in the United States and around the world.[9] The ERS are observational tools that capture in standardized formats information on a host of features in the settings that serve young children, including physical safety, hygiene, nutrition, educational materials, program offerings (for example, activity schedules), and qualities of social and language interactions between adults and children. Observers are trained for agreement with master-coded examples and demonstrate specific levels of accuracy before using the system in the field. A combination of observation and interviews are used to gather data, all of which yield quantitative scores for program features plus an overall global scale for quality. The Early Childhood Rating Scale, or ECERS, is one of a suite of environmental rating scales, or ERS, for children from birth to five years old. There are ERS for infants, toddlers, and for family child care.

ECERS is the most widely used metric for program quality in early childhood education settings such as Head Start, preschool, and subsidized child care.

It would be difficult to overstate the importance of the environmental rating scales, particularly the ECERS, in early childhood education program development and policy. Nearly every single public investment in early childhood education—from increasing access or slots in existing programs to opening new sectors of programming to improving existing programming—has involved legislative or regulatory language related to ensuring quality. For more than three decades, the ERS have been the gold standard.

The ECERS has had a ubiquitous presence in most major studies of early education quality and impacts, including national-level evaluations of Head Start and Early Head Start program quality and impacts.[10] The scales have been used in studies and program-improvement efforts in Canada, most European countries, and increasingly in Asia. In each use the scales have proven reliable and valid and required only minor adaptations in each country. Nearly all of these studies used large and diverse samples of children, teachers, and settings. These research studies not only provided data on the validity and use of these rating scales, but also considerable experience in the development and deployment of regimes for training, quality control, and scoring. Because the ERS were designed to capture properties of settings and adult-child interaction thought to be relatively invariant across the range of U.S. settings—family day care, private preschools, Pre-K, and Head Start—perhaps it is not surprising to find that these features operate similarly in other western industrialized countries.

It would be difficult to overstate the importance of the environmental rating scales, particularly the ECERS, in early childhood education program development and policy.

Nearly all the research on ERS over the course of the 1980s, 1990s, and into the early 2000s, finds a relation between higher scores on the ECERS and more positive child development outcomes in areas that are considered important for later school success, such as language development.[11] Of interest is that more recent studies of state-funded, prekindergarten and Head Start programs have found fewer and more modest associations between ECERS scores and children's growth on school-readiness assessments, a pattern that will be explored in greater detail later in this paper.

As noted earlier, environmental rating scales are used in a variety of ways, including high-stakes applications as well as for self-assessment by center staff, preparation for accreditation, and voluntary improvement efforts by licensing or other agencies. More than 20 states use ECERS as one of the metrics on their Quality

Rating and Improvement Systems, or QRIS,[12] an accountability and program-development policy tool that figures prominently in the recent federal investment in early childhood education, specifically the Early Learning Challenge grants that are part of Race to the Top. In most QRIS models several metrics hypothesized to be part of program quality (for example, quality of the environment, teacher credentials, features of the curriculum to name a few) are combined to derive an overall rating of quality (for example, three stars in a five-star rating system) that can serves as a signal to improve quality. States are investing in program improvements and professional development that are purportedly coupled with QRIS metrics. Although states' algorithms for combining quality metrics and the specific quality metrics themselves vary, the ECERS is featured in most.[13]

Subsequently, there are an abundance of examples of scaled-up use of standardized observation using the ECERS that align with policy initiatives and program-development investments in quality improvements. Overall these efforts affect millions of children.[14] Evident throughout all these uses is how standardized observation is a fundamental component of systems that serve both an accountability aim (for example, tiered reimbursement for services contingent on observation metrics, a policy innovation that could apply in K-12 for something like Title I programs or special education) and program-improvement aims (for example, coaching or investments in credentialing). Features of early childhood programs specified on the ECERS indicators are also woven into professional licensure and credentialing systems. This is an example of observational indicators linking back into professional-preparation program content and the systems that credential professionals and license settings. Several states offer certificates through which early childhood professionals receive credit, licenses, and program accreditation based directly on their production of items on the ERS.[15]

As previously noted, the ERS, particularly the early childhood environment rating scale, have been a policy target for accountability and improvement. Public investments in early childhood have been linked in policy or regulation to raising ECERS scores and have gone directly to the features of programs and settings assessed by the ECERS. This linkage demonstrates very clearly that even for observational assessments, metrics that have stakes attached tend to change over time, in other words, what gets measured gets done. With more than two decades of investments in Head Start, ECERS scores gradually increased nationwide to the point that the mean score in nationally representative reports showed an overall quality level of "5" on the ECERS seven-point range.[16] Features of quality measured by the ERS that include materials, the physical environment, hygiene, or program schedules have primarily accounted for the reported jumps in scores. These increases have undoubtedly improved the experiences of children, the safety of settings, and the overall quality of programs. Further, in several cases these improvements appear to also have corresponded to improvements in some measured aspects of children's development.[17]

Yet other features of programs measured by the ERS, including aspects of adult-child interactions, have been much harder to improve. Moreover, recent studies, including those tracking Head Start, show that ERS-defined quality improvements have not directly led to improvements in children's school readiness. To the extent that the features of early childhood programs assessed by ECERS show considerable variation, then the use of ECERS in these large-scale program improvement and accountability efforts was associated with incremental increases in child outcomes. When programs lack educational materials or fail to operate with a daily schedule of learning activities (indicators on the ECERS), then a focus on those benchmarks translates into increments in children's outcomes. But when nearly all programs get "up to speed" on ECERS-defined quality and variation in those features declined (such as occurred in Head Start), links between programs' ECERS scores and child outcomes also appeared less strong. Further analysis of these patterns of results related to quality assessment and improvement revealed that other elements of observed program quality (for example, teacher-child interactions) were potential candidates for more focused assessment. In some sense there was evidence of an accountability-framed observational assessment pushing improvement to the point that there was a ceiling effect on the assessment.

In a very real way, these examples show how observation can be embedded into accountability and improvement models such as those being discussed presently in K-12 and actually drive change in observed indicators. In short, experience with the ERS protocols in a wide range of large-scale deployments indicates that observations can be scaled and used in accountability, program development, and market-oriented policy tools to produce, over time, change in those features of programs assessed by those tools.

CLASS—Classroom Assessment Scoring System

The Classroom Assessment Scoring System, or CLASS,[18] is a more recently developed observational instrument designed to measure features of teacher-child interaction in settings serving children as young as infancy and extending, with different versions, through high school. Currently, however, the CLASS has been most widely used in preschool classrooms.[19]

The CLASS dimensions are based on development theory and research suggesting that interactions between children and adults are a primary mechanism of development and learning, a tenet widely held to be the case for younger children and recently validated for students in middle and secondary grades as well. Unlike the ERS observation system, the CLASS metrics focus only on interactions between teachers and children in classrooms (scoring for any dimension is not determined by the presence of materials, the physical environment, safety, or the adoption of a specific curriculum). This distinction between observed interactions and physical materials or reported use of curriculum is important because in most early elementary settings materials and curriculum are usually prevalent and well organized. With the CLASS the focus is on what teachers do with the materials they have and the interactions they have with students. In addition, it complements the information gathered by the ECERS.

The CLASS is aligned with a set of professional development supports such that teachers are helped to make positive changes in the areas of their practice with which they struggle.

Importantly, the scoring guides, manuals, training materials, and initial validity testing for the CLASS were developed through use in two large-scale national studies involving observations of early education classrooms—the National Institute of Child Health and Human Development study of early care and youth development[20] and the National Center for Early Development and Learning Multi-State PreK Study.[21] These studies provided a wealth of experience and information on scaling up standardized classroom observations of teacher-child interactions in more than 5,000 Pre-K–fifth grade classrooms and created a strong research and evidence base for a host of practical decisions and resources.

The CLASS describes three broad domains of teachers' interactions with children—emotional support, classroom organization, and instructional support—that are common across teacher-child interactions from preschool to 12th grade. Within each domain there are several specific dimensions of interaction that vary by grade. The CLASS measures effective teacher-student interactions across Pre-K-12 in a way that is sensitive to important developmental and context shifts that occur as students mature. The CLASS is aligned with a set of professional development supports such that teachers are helped to make positive changes in the areas of their practice with which they struggle.

The CLASS, like the ECERS, is widely used in research and program development as well as in Head Start and QRIS systems. These uses require standardized training and reliability testing protocols. In the past three years more than 4,000 people across the country have been trained to reliably use the CLASS—thus documenting its scalability. As with the ECERS, there are a variety of training opportunities that allow districts and states to effectively use the CLASS on a large scale, including a fully developed and tested train-the-trainer model. Most of the CLASS observation training takes place in face-to-face training workshops following trainees' completion of a set of preparation assignments and video review that can be done on the web. The most recent versions of the CLASS, developed for use in upper elementary and secondary classrooms, rely extensively on the web as the mechanism to support training to acceptable levels of reliability.

It is evident from the work done on training with the CLASS and with the ERS, that large- scale, national-level implementation and rollout of an observational assessment is possible with combinations of live and web-based training protocols to sustain the training of thousands of observers to acceptable levels. A growing body of work now documents the ways in which the CLASS observations from Pre-K-12 settings identify components of teacher-student interactions that contribute to students' social and academic development.[22] The pattern of results is quite clear: teachers' instructional support (feedback, focus on conceptual understanding, rich conversational discourse) are overall low; at the same time, instructional support behaviors appear to be strong predictors of students' learning gains. Importantly, it has also been demonstrated that these teacher instructional behaviors can be improved by professional development.[23]

The CLASS is also used in a variety of high-stakes and program-improvement applications. In recent federal legislation reauthorizing Head Start, it was specifically mentioned that a standardized observation of teacher-child interaction was to be the metric for program monitoring and accountability. The CLASS was chosen as this measure and in the spring of 2009 large-scale training and train-the-trainer workshops were launched to achieve a national rollout. As an analogue to the use of observations in K-12 accountability systems, every Head Start grantee (grantees range in size from a few to many hundred classrooms and are the fiscal unit of allocation) is evaluated every three years with CLASS observations conducted in a representative number of classrooms by a set of independent, trained evaluators. Cutoff scores have been established based on the accumulated empirical evidence on the CLASS that designate levels of scores that are acceptable for continued operation of a Head Start program. In effect, observations will be used as a component of measuring Head Start grantees' performance: If classrooms are not meeting certain standards for qualities of teacher-child interactions then a grantee will have to compete again for Head Start funding.

In parallel to this accountability-driven evaluation use, the Office of Head Start has funded a network of training and technical-assistance centers, early childhood specialists, and related personnel to focus on program improvements and human-capital advancement, much of which focuses on the CLASS and associated professional-development programs that have been demonstrated to improve the CLASS scores. It is estimated that as many as 25 percent of current Head Start grantees could fall below the CLASS cutoffs for quality and would therefore have to reapply on a competitive basis for Head Start funding.

Like ECERS, the CLASS is also being used in Quality Rating and Improvement System models for preschool and child care programs in a variety of states. New Mexico, Florida, Georgia, Massachusetts, Pennsylvania, and others have adopted the CLASS as one of their QRIS metrics. In fact, several states are using both the CLASS and ECERS in their QRIS models, thus relying heavily on standardized observation for accountability and program improvement.

It is too early to tell the extent to which high-stakes adoption of the CLASS in early childhood-accountability or program-improvement systems has resulted in an actual shift in program quality or in children's school readiness. It is, however, quite evident that the system's use in this framework has driven grantee's attention and requests for training and technical assistance to the degree that early childhood education is now very focused on teachers' instructional interactions. Clearly, between the ECERS and the CLASS, early childhood education has accumulated a wealth of experience in using standardized observations in policy and program-improvement contexts and in deploying observational protocols. It is this experience and the base of information garnered from research studies and evaluation that provide the basis for the lessons learned that we examine next.

Three key considerations when using observation in large-scale applications

Research and experience with using observation in large-scale applications (districts, states, nationwide) in early childhood education programs has enabled the accumulation of evidence in three key areas related to using classroom observations. These three areas are:

- Reasons to observe classrooms and teachers—we present a model for understanding how observing teachers' behaviors plays an important role in organizations geared toward systematically producing higher quality opportunities for classroom learning. This includes research-based information on several key areas of teachers' observable practice and how those practices impact learning.

- Choosing and using observation tools—we outline key questions that can guide instrument selection that are aligned with strategic program goals. We also include a list of guiding principles for the successful use of observation tools, as well as logistic information regarding important ways to standardize observation protocols.

- Using data from observations to systematically improve the quality of classroom practice—we review strategies for translating observational findings into effective feedback for teachers and offer guidelines for presenting observational findings to teachers in ways that support them in making practical shifts to maximize student growth and development.

Reasons to observe classrooms and teachers

Teaching and learning is a system where teachers' behavior and instruction are embedded in and influenced by supports and constraints that are important to consider. In order to understand why and how standardized, valid classroom observations can improve student outcomes, it is helpful to see how these observations are embedded within an overarching framework for recognizing how learning and development take place for both teachers and students.

Specifically, we see three key and linked aspects of the teaching-learning system which are represented in Figure 1:

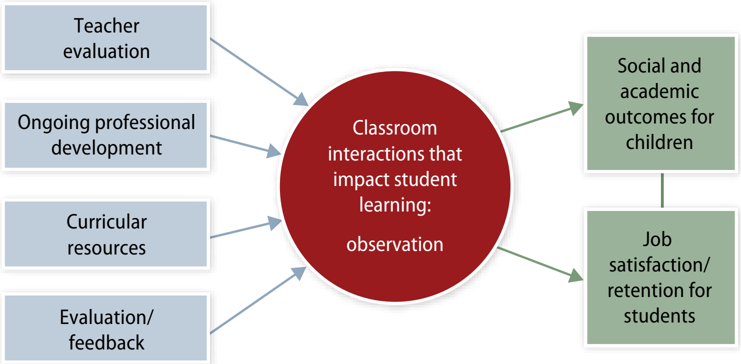
- Inputs/resources
- Teachers' interactions with children
- Outcomes such as student learning

Starting with inputs, we looked to literature in the fields of adult learning and professional development (in education as well as in other fields) to better understand the resources that support the acquisition of a set of behavioral competencies in teachers, which translate into improved learning outcomes for students. We found four areas that seemed key to helping teachers develop these competencies: providing teachers with knowledge about effective practices; providing professional development that is individualized, classroom practice-based, and ongoing; providing curricular resources and materials; and providing specific feedback on teachers' own practice.

The skills that teachers develop as a result of these inputs can foster effective interactions with students. Observations of teachers' interactions and classroom processes play a major role in helping describe and identify effective practices and improving these practices through professional development. Thus observation can be an effective tool in building capacity for teaching and learning.[24]

FIGURE 1

Links between Inputs and Outputs


Teacher evaluation
Ongoing professional development
Curricular resources
Evaluation/ feedback
Classroom interactions that impact student learning:
observation
Social and academic outcomes for children
Job satisfaction/ retention for students

Observing teachers' classroom interactions and practices is one element of assessing how this instructional system is operating and a potentially key lever for improvement. It is not the only element, however, of the system supporting children's learning. To make the point, consider that in many early childhood classrooms teachers exhibit qualities of interactions with students that are consistent with children's learning gains, but in the absence of curricula that can focus those interactions on key skills and knowledge, little learning actually occurs. This is particularly true in areas in which curriculum is underdeveloped, such as math or science. Relatedly, many elementary school teachers exhibit positive features of interaction and instruction but lack of knowledge in a particular content domain (for example, math or science), undermining the impact of those interactions on student learning. The use of standardized observations, if they reliably and validly measure classroom interactions that impact student learning, is a direct and effective mechanism for focusing on teachers' classroom interactions with the potential to illuminate links between certain inputs (resources for teachers) with desired outcomes (optimized student learning).

Certainly this is not a new or novel idea. Every principal spends time observing teachers and most teacher-education programs have some way of providing future teachers with feedback on their practicum experiences in classrooms. Still the vast majority of these observations rely on unstandardized, informal, and nonvalidated procedures. Each school district, principal, and mentor-teacher derives their own set of ideal teacher practices, some based on empirical research and some simply a reflection of personal preference or broad educational theory. Without the more systematic use of standardized, reliable, and validated observational tools, the ultimate value of these observations and the feedback they provide to teachers is limited, particularly when the aims of such approaches include documentation and improvement of practices in a very large number of classrooms (often in the thousands). Without a standardized, validated system in place, teachers are likely to receive very different types of feedback and support depending on grade-level, school or on the person doing the observing. Such approaches are unlikely to build capacity in a school or district nor result in system-level improvements over time.

Without a standardized, validated system in place, teachers are likely to receive very different types of feedback and support depending on grade-level, school or on the person doing the observing.

The advantage of using tools that are standardized, reliable, and validated against student outcomes is that educators, mentors, and administrators can make comparisons on an even playing field. When noting strengths and challenges across classrooms, observers can see and note behaviors directly related to student growth and development.[25] The use of these tools in no way interferes with giving personalized feedback to teachers. Instead it allows for highly specific and individualized feedback with regard to clearly defined areas consistent across all teachers, while also providing a strong background for interpretation of scores. Further use of standardized tools outweighs the disadvantages related to a highly customized approach in which every classroom, school, or district adapts an existing tool or develops a new one, particularly because these type of customizations rarely if ever have the strong technical properties (reliability, validity) of existing tools. As a consequence the resulting hybrids often cannot support the desired interpretations and uses (for example, tenure decisions, inferences about improvements, and more).

We next discuss these specific features of observational protocols—standardization, reliability, validity, link to professional development—and the role they play in the selecting an observational system.

Choosing and using an observational system

In the swirl of competing interests—teachers' unions, teachers, reformers—school district leaders find themselves wanting and needing to act and having to make difficult decisions. In this context deciding to use observations of teachers as a component of performance assessment is perhaps the least complex decision school leaders face. Still there are a host of questions and concerns that go into choosing a particular observational system and the procedures involved in implementing that or any observational approach.

In this section we describe:

- The focus of an observation and the nature and scope of behaviors observed
- Standardization of protocols and procedures; reliability and training
- The validity of observations as measures of teacher or classroom qualities
- Additional complementary supports for implementation and use

In each of these areas, lessons learned from large-scale use observations in early childhood settings are presented along with vignettes that present actual applications and situations that translate these lessons into actions and decisions in K-12 schools.

What teaching practices do observational tools assess?

There are multiple published and unpublished classroom observation systems available for use and deciding among them is the first step in putting an observational system to work.[26] The primary advantage of using an existing observation tool is that it saves a great deal of time and resources that would otherwise be put into developing a new instrument, even one with minimal levels of reliability and validity for predicting outcomes of interest.

Different instruments provide users with different types of information about classrooms. Some are quite broad in nature, providing data on the physical environment, the types of activities, or the teacher's execution of professional responsibilities such as record keeping and communicating with families. Others adopt a more focused approach, such as exclusively attending to a specific set of instructional interactions that take place within short observation windows or focusing on comparisons between the experiences of specific groups of students within the classroom. Still others strike a balance in terms of scope, including information on a variety of teacher and student behaviors but excluding information that would require knowledge outside of what is obtained during specified observation windows (for example, not including information about how a teacher communicates with parents, makes lesson plans, and more). It is important that users begin by defining the goals that their organization has in using a particular observation tool. After defining the desired output information, users can then select a measurement tool that is aligned with their objectives.

In addition to ensuring a match between the scope of an observation instrument and the defined goals of an organization, users are advised to consider the specific design of the instrument, including its age range and the grade levels from which data on the psychometric properties of the instrument have been obtained. If your goal is to assess fourth-grade classrooms, for example, it is ideal to use an instrument that was generated with this developmental level in mind and has been validated for use with this age group.

Relatedly, some users may want to focus more on the provision of general support for learning, whereas others may have programmatic goals that focus more specifically on the quality of instruction in different content areas, such as mathematics or reading. There are instruments available that assess implementation of content-specific learning supports as well as tools that focus on supports linked to student growth and development across content areas. If an organization has a particular interest in a certain content area, they may wish to supplement a protocol for observing generalized supports with one that includes specific interactive practices relevant to the content area of focus.

Focusing observational protocols

Content specific or more general?

The fictional Fairmont school district is considering mandating the use of a new mathematics curriculum in all of its schools. A small number of teachers who are pilot testing the new curriculum have been trained on this approach to teaching mathematics and have been provided with all needed materials. The district is now looking to evaluate the extent to which teachers using the new curriculum are incorporating high-quality strategies for teaching mathematics in comparison with the extent to which teachers in a control group of schools are also incorporating such strategies in their math classes. The aim of the evaluation is to help the district decide whether the new curriculum is a good choice for districtwide use.

In this scenario the Fairmont school district may wish to use an observation protocol that is focused on research-based definitions and descriptions of high-quality mathematics instruction or to supplement a more generalized observational protocol with a content-specific protocol for mathematics instruction.

In contrast to Fairmont, the make-believe Lakeview school district wants to conduct an observational assessment of all its teachers in order to gain a better understanding of systemwide areas of strength and weakness that will enable the district to plan for in-service programming and create individualized professional-development plans for teachers. Observers will conduct multiple observations per day, which means these observations will occur at different times of day and during different activities for different teachers.

The Lakeview district would likely benefit from use of a protocol that is designed to assess generalized supports for learning that produce benefits for student development across content areas since not all teachers will be observed teaching the same content areas.

An additional consideration that falls within this question concerns the specificity, or "granularity," of the behaviors being observed. For example, is the observational system capturing information on specific, highly discrete teacher behaviors (for example, counting the times the teacher praises a child) or on more global, but well-defined patterns of behavior that unfold over a lesson or period of time (for example, a tendency to use a variety of ways to motivate students)? Measures using *frequency counts or time-sampling methodology* ask users to count the number of specific types of behaviors observed in a specified time window (usually short in length). *Global ratings* guide users to watch for patterns of behavior and make integrative, summary judgments about value, nature, or quality of those behavioral patterns. Some examples of behaviors assessed by time sampling measures include time spent on literacy instruction, the number of times teachers ask questions during instructional conversations, and the number of negative comments made by peers to one another. In contrast, global-rating systems may assess the degree to which literacy instruction in a classroom matches a description of evidence-based practices, the extent to which instructional conversations stimulate children's higher-order thinking skills, or the extent to which classroom interactions contain a degree of emotional and behavioral negativity between teachers and students and among peers.

Recalling the earlier discussion about the early childhood environmental rating scale and how program-quality investments tracked the metric, particularly the features of programs that reflected materials and the physical environment, the lesson there was that observational indicators drove investment and training in ways that changed levels on those indicators. Specificity of the actual observational indicator matters here. To the extent that what gets observed gets done, then observational approaches that focus on counting behaviors (for example, the number of open-ended questions a teacher asks or the frequency with which a teacher does a specific action) will drive increases in those discrete behaviors as the observation rolls out into accountability of program improvement work. There is a tradeoff with specificity, however. Generally speaking, it is easier to obtain high levels of reliability for highly specific and discrete behaviors using counting or time-sampling collection methods. But those discrete indicators have shown little power in relation to predicting student learning gains. Rather, data collected over time that capture broader yet well-defined features or patterns of interaction tend to be better contextualized to the individual classroom setting and better demonstrate predictive power in relation to accounting for student learning. More general codes focused on patterns of interactions and behaviors require some judgment by observers and hence are more challenging with regard to reliability and training while showing stronger relations with student learning.[27]

There are advantages and disadvantages to each type of system. An advantage of global ratings is that they assess how behaviors are organized and results can be more meaningful to teachers rather than a simple count of discrete behaviors in isolation. To illustrate this point consider the act of smiling by a teacher, which can be termed a teacher's positive affect. This act of smiling can have different meanings and may be interpreted differently depending on the response of students in the classroom. In some classrooms teachers are exceptionally cheerful but their emotional displays are inconsistent with those of students. Other teachers are more subdued in their emotions but there is a clear match between teacher and student experience. A measure that simply counted the number of times a teacher smiled at students would miss these more nuanced interpretations. In this case an observational instrument, with a focus on frequencies of specific behaviors may lend itself to easy alignment with the evaluation of focused interventions. If a goal is, for example, to increase the numbers of times teachers provide students with specific feedback, then time-sampling methods could be useful. Time sampling could yield specific data on intervention effects on feedback by counting the frequencies of specific feedback behaviors before and after the intervention (or in classrooms that did and did not receive the intervention). Similarly, the success of an intervention designed to increase the amount of time spent in learning activities (versus "down time") could be evaluated using time sampling methods.

One other difference related to the granularity of observations concerns the degree to which specificity is related to observer effects. Scores obtained from global ratings appear to contain more information about the observer than time-samplings of more discrete behaviors. This finding is not surprising given that global ratings tend to require greater levels of inference than do frequency approaches. Counting the number of times a teacher smiles, for example, requires much less inference than does making a holistic judgment about the degree to which a teacher fostered a positive classroom climate. This point emphasizes the need for adequate training and strategies for maintaining reliability among classroom observers, issues we consider in greater detail shortly.

The apparent advantages of more discrete behaviors in terms of somewhat lower observer-related variance, however, are counteracted by a number of other facets of observation. This brings us to another factor to consider: the extent to which an observational score can be attributed to stable characteristics of a teacher versus factors that change over time as a result of a number of variables, including subject matter, number of students, and time of day. This is a very important consideration when the desired outcome of the observation is to make some inference about a teacher's skills or capacity. Evidence clearly suggests that more discrete, specific behaviors such as those that can be counted or time sampled do not capture stable features of teachers or classrooms, whereas more global ratings that capture patterns of behavior reflect properties of a specific teachers' approach to interaction that remain stable across periods of the day, days of the week, months, and even content areas. Highly specific and discrete codes do not appear to capture the behavioral tendencies of teachers that are stable across time or that distinguish between different teachers' styles.

Is the observation protocol standardized in terms of administration procedures and does it offer clear directions for conducting observations and assigning scores?

It is important to select an observation system that provides clear instructions for use, both in terms of how to set up and conduct observations and how to assign scores. Without standardized directions to follow, different people are likely to use different methods, which severely limits the potential for agreement between observers when making ratings, thus hampering systemwide applicability.[28] In this regard standardization is not the same as reliable or valid, instead it refers to the rules and procedures for observing and ensuring consistency and quality control in how information is collected. These procedures include considerations of time of day, qualifications of observers, length of the observation, and other features that could undermine the quality of data collected and ultimately the inferences drawn from those data.

Importance of standardization for observational instruments

A teacher-preparation program is looking for a way to assess their students' performance at the beginning and end of their student-teaching experience, during which time they are also taking a course on effective teaching practice. Program officials find "Observational Protocol A," which has six clearly defined, theoretically based, 10-point scales that observers use to rate teacher practice. Several members of the faculty read the definition of the six scales and agree that the teaching behaviors the scale assesses are aligned with the course objectives as well as with the broader goals of the program. It is decided that the six scales would be good targets for assessment. The program selected, however, does not include training or observational protocols or explicit directions for scoring. As a consequence, Observational Protocol A is used quite differently by the two faculty members in assessing student performances.

When Professor A makes observations he arranges the observation time in advance with the teachers. He arrives at the appointed time, but does not begin the observation until he can tell that the teacher is ready to begin the lesson and he ends the observation as the teacher ends the lesson. During this time he takes detailed notes about the teacher's practice along the six dimensions. When scoring, he reasons that if he sees a teacher engaging in the behaviors under consideration several times, they should get "full credit," or a 10, on the scale.

Meanwhile, Professor B also conducts observations using the same well-defined scales, but her visits are unannounced. She typically arrives at the beginning of the school day and begins taking notes as soon as she arrives and observes for two consecutive hours, regardless of start and stop time of activities. In terms of scoring, she reasons that teachers start at a "1" level and she moves the score up a point on the scale every time the teacher successfully engages in the behavior under consideration. Given these differences in protocol, it is likely that Professor A's scores could be systematically higher than Professor B's.

continued on next page

This example shows that even with well-defined codes, it is extremely important to have a clear observation and scoring protocol that all observers follow in order to obtain scores that are consistent across observers. In this example, note that significantly different scores are likely to result from Professor A's observations and Professor B's observations as a result of their different administration and scoring techniques, and that these scores may or may not reflect real differences between the two teachers they observed. For instance, if Professor A used his interpretation of the protocol to conduct initial start-of-student-teaching observations and Professor B used her interpretation of protocol to conduct the end-of-student-teaching observations, any real gains in teaching practice could be obscured. What's more, the preparation program might conclude that the course and teaching experience did not function as effective preparation when in fact, if the teachers were evaluated using the same protocol on both measurement occasions, they might have shown improvements.

There are three main components of standardization that users may consider when evaluating an observation instrument: training protocol, parameters around observation, and scoring directions. With regard to the training protocol there are several questions: Are there specific directions for learning to use the instrument? Is there a comprehensive training manual or user's guide? Are there videos or transcripts with gold standard scores available that allow for scoring practice? Are there other procedures in place that allow for reliability checks such as having all or a portion of observers rate the same classroom (live, via video, or via transcript) to ensure that their scoring is consistent? Are there guidelines around training to be completed before using the tool such as do all observers need to pass a reliability test, observe in a certain number of classrooms, or be consistent with colleagues at a certain level?

Regarding parameters around observation, users are also advised to look for direction and standardization in terms of the length of observations, the start and stop times of observations (are there predetermined times, times connected with start and end times of lessons/activities, or some other mechanism for determining when to begin and end?), time of day, specific activities to observe, whether observations are announced or unannounced, and other related issues.

As for scoring, users are advised to look for clear guidelines. Some questions to consider: Do users score during the observation itself or after the observation? Is there a predefined observe/score interval? How are scores assigned? Is there a rubric that guides users in matching what they observe with specific scores or categories of scores such as high, moderate or low? Are there examples of the kinds of practices that would correspond to different scores? Are scores assigned based on behavior counts or qualitative judgments? How are summative scores created and reported back to teachers?

Does the observation include reliability information and training criteria?

Reliability is a key consideration in selecting an observational assessment tool.[29] Reliability is a property of any measurement tool that refers to the degree of error or bias in the scores obtained. It addresses the extent to which a tool measures those qualities consistently across a wide range of considerations that could affect a score, for example, the raters themselves, the length of the observation period, and observer training. In observational assessments of classrooms, a reliable tool produces the same score for the same observed behaviors regardless of features of the classroom outside of the scope of the tool and regardless of who is making the ratings. Just as a yardstick registers the same number of inches when measuring a given sheet of paper, regardless of whether that paper is measured during the day or at night, inside or outside, or who is holding the yardstick, a tool that measures teachers' ability to promote student language should produce the same scores for the same behaviors, regardless of whether these behaviors occur during math or literacy, whole group or small group, and regardless of who is making the ratings.

Consistency is the foundation of observation

Let's consider the experience of two observers who we will call Principal Menendez and Vice Principal Edwards. Both individuals are conducting observations in their school using the same standardized protocol on which they have both been well trained. Menendez and Edwards both want to make sure that they are consistent not only with the scoring manual, but also with one another since they will split classrooms between them and do not want differences between the two of them to result in unfair advantages or disadvantages in the ratings the classrooms are given. Therefore, they decide that on a regular basis, once every 10 observations, for example, they will go into classrooms together, observing and rating the same lesson to check the consistency of their scores. They frequently find that they are scoring reliably, however, if there are discrepancies between their scores, they discuss them to make sure that they are interpreting behaviors consistently with the instructions supplied by the system. They find that this keeps them from drifting from the scoring protocol outlined in the manual and gives them confidence that they are truly using the same yardstick to measure the performance of all teachers in their school, regardless of who is conducting the observation.

In another example, observer Brown and observer Yang both conduct classroom observations assessing the efficacy of teachers' behavior-management techniques among other things. Observer Brown is rating a classroom in which a teacher is working with a group of 10 students on a hands-on science lesson. The teacher engages in effective behavior-management techniques, her expectations are clear, and she helps the students learn to regulate their own behaviors in positive, efficient ways.

Meanwhile, observer Yang is rating a different classroom in which the teacher is managing the behavior of a group of 23 students as they wait for a guest speaker who is unexpectedly delayed. This teacher engages in the same kinds of behavior-management techniques as in the science classroom—expectations are clear, the teacher is positive and effective, and helps the students learn to self-regulate their behaviors. Despite the differences in group size and classroom activity, these two teachers receive the same scores on the behavior management scale because they are engaging in the same types of behaviors with the same levels of efficacy. These two teachers may receive different scores in other areas such as questioning or use of time, but their behavior-management techniques were equivalent in quality and thus are scored the same.

There are several aspects of reliability, but perhaps the two most relevant when considering classroom observation systems are *stability over time* and *consistency across observers*.

Turning first to stability over time, assuming a goal is to detect consistent and stable patterns of teachers' behaviors, users need to know that constructs being assessed represent a stable characteristic of the teacher across situations in the classroom and are not random occurrences or behaviors that are linked exclusively to the particular moment of observation. If ratings shift dramatically and randomly from one observation cycle or day or week to the next, these ratings are not likely to represent core aspects of teachers' practice. Conversely, if scores are at least moderately consistent across time, they likely represent something stable about the set of skills that teachers bring into the classroom setting and as a result feedback and support around these behaviors is much more likely to resonate with teachers and function as useful levers for helping them change their practice. It is advantageous if observational tools provide information on their test-retest reliability or the extent to which ratings on the tool are consistent across different periods of time (within a day, across days, across weeks, or more).

There are several aspects of reliability, but perhaps the two most relevant when considering classroom observation systems are *stability over time* and *consistency across observers.*

A notable exception around the criteria of stability over time as a marker for reliability, however, is when teachers are engaged in professional-development activities or are otherwise making intentional efforts to shift their practice. In these cases, as well as in cases where a school or district's curriculum is changing or new programwide goals are being implemented, a lack of stability in observations of teacher behaviors may well represent true changes in core characteristics and not just random (undesired) fluctuation over time. In these cases it would be desirable to collect data on the extent of change and specific areas where change is observed.

With regard to stability across observers, in order for results of observations to be useful and valid, training protocols and provisions of scoring directions must be clear enough to produce agreement across observers. If there is very low agreement between two or more observers' ratings of the same observation period, the degree to which the ratings represent the teachers' behavior rather than the observers' subjective interpretations of that behavior or personal preferences is questionable. Conversely, if two independent observers can consistently assign the same ratings to the same patterns of observed behaviors, this speaks to the fact that ratings truly represent attributes of the teacher as defined by the scoring system as opposed to attributes of the observer. Therefore, users may wish to select systems in which there is documented consensus among trained raters to what extent teachers are engaging in the various behaviors under consideration.

If there will be several different observers making ratings, an important consideration is how much variability in scores can be attributed to the raters themselves.[30] Not surprisingly, rater effects are significantly higher when using observation systems requiring raters to make global judgments than with time-sampling systems that provide counts of low-inference behaviors. Almost every observational system, however, will have some rater effects and therefore it is important to be aware of these effects and make efforts to keep them to a minimum regardless of the type of observation system being used.

If there will be several different observers making ratings, an important consideration is how much variability in scores can be attributed to the raters themselves.

Rater effects are most relevant if there will be multiple people conducting observations within a given system. Even if a single individual is conducting all observations within a school, and if these ratings will not be used in comparison to ratings completed by other raters or in other schools, it is still important for each observer to receive excellent training on the instrument, meet "gold-standard" criteria prior to conducting observations, and to take periodic "drift" tests to ensure that they remain reliable with the standards outlined by the developers of the measure such as those standards that have proven links to student outcomes. When there are several different observers, the importance of this issue is multiplied as each individual observer must maintain reliability with both the "gold-standard" criteria of the instrument developers as well as with one another.

Several steps can be taken to minimize rater bias.[31] First, it is important to select tools that are well standardized and have documented potential for reliable use across observers. In addition, implementing a high-quality training program for all observers will help ensure that raters are more consistent with one another. Similarly, including periodic "drift" testing at predetermined intervals (annually or biannually if observations are conducted for professional-development purposes and monthly if data will be used for accountability purposes) can offer a refresher in scoring procedures and help improve the degree to which raters remain consistent with scoring protocols and with each other.

With regard to scheduling observations/assigning raters to classrooms, rotating raters across teachers can help avoid systematic variance in scores. If, for example, all classrooms are visited twice over the course of the year and Vice Principal Smith and curriculum coordinator Jones share observation responsibilities, consider having each rater observe each classroom one time. Random assignment of observers to classrooms can also be useful in reducing systematic rater bias. Alternately, if time and resources allow, multiple raters can observe and rate classrooms simultaneously and their scores can be averaged thereby reducing the amount of bias introduced by any single observer.

Is there evidence for the validity of the observational metrics?

Validity represents the degree to which scores or metrics derived from the observation system are associated with specific student or teacher outcomes. Along with reliability considerations, validity is one of the most important aspects to consider when selecting an observation instrument. Different observation systems have varying levels of data available on how closely aligned the outputs of observations are with students' performance in a specified area, students' growth on specified skill sets or other outcomes of interest.

Selecting instruments with demonstrated validity is critical to making good use of observational methodology because this information allows users to have confidence that the information being gathered is relevant to the outcomes that they are interested in and that the types of behaviors outlined in the system can be held up as goals for high-quality teacher practice. Without validity information users have no such assurances. Knowing that assessment tools are directly and meaningfully related to outcomes of interest before they are used either in professional development or accountability frameworks is important.

Equally important is clarity. A system may be valid for one set of outcomes but not for another, so clarity around outcomes of interest is key. An observation system, for example, may include validity data regarding the prediction of students' academic achievement during that school year, but it may demonstrate no relation to student dropout rates in subsequent years. If the objective of conducting the observation is to evaluate whether teachers are engaging in behaviors that promote students' learning over the course of the year, this may be a well-suited instrument for that purpose. But if the objective is to determine whether teachers are enacting behaviors that will prevent students from dropping out, a different observation with documented links to dropout rates may be preferable.

If a user has a particular observation tool that is aligned with the questions they want answered about classroom practice and meets the criteria summarized previously (for example, standardized, reliable), there is always the possibility that no data will be available on validity for the particular outcomes that the user is interested in evaluating. In these instances, it would certainly be possible to use the observation in a preliminary way and evaluate whether it is, in fact, associated with outcomes of interest. A district, for example, could conduct a pilot test with a subgroup of teachers and students to determine whether scores assigned using the observation tool are associated with the outcomes of interest. This testing would provide some basis for using the instrument for accountability or evaluative purposes.

In sum, the importance of selecting an observation system that includes validity information cannot be overstated. It may be difficult to find instruments that have been validated for your purposes, but this is truly essential for making observational methodology a useful part of teacher evaluation and support programs. If the teacher behaviors that are evaluated in an observation are known to be linked with desired student outcomes, teachers will be more willing to reflect on these behaviors and "buy in" to observationally based feedback. Further, teacher educators and school personnel can feel confident establishing observationally based standards and mechanisms for meeting those standards, which means educational systems, teachers, and students will all benefit.[32]

The importance of complementary sources of information

Obtaining information about classrooms from multiple sources and from different perspectives, including the perspectives of teachers, students, and individuals who are generally familiar with the classroom on a routine basis, as well as the observers' data collected during the specific observation window, can provide a more comprehensive picture of the classroom environment. This can also be helpful in terms of providing constructive feedback in that one could seek out coherent patterns in responses across observers/raters. Having a teacher engage in a self-study or self-assessment in conjunction with structured observations made by neutral observers may be an example of a useful way of facilitating goal setting and problem solving with teachers. Likewise, obtaining students' perspectives can be an invaluable resource in understanding how specific teacher behaviors impact students' subjective experiences of the classroom. Equipped with this information, those providing feedback to teachers may be able to present a richer picture of what is happening in the classroom and how that impacts all classroom participants, including the teacher's own feelings of efficacy and students' experiences of support and challenge in the classroom.

As the goals of conducting observations include not only gathering information on the quality of classroom processes but also using that information to help teachers improve their practices (and, eventually, student outcomes), observation systems that include a protocol to assist in translating observation data into professional-development planning is desirable. Information such as national norms and threshold scores defining "good enough" levels of practice (levels of quality that result in student improvement), or expected improvements in response to intervention would be extremely useful to have, although few, if any, instruments currently provide this kind of information to users.

Also useful are guidelines or frameworks for reviewing results with teachers, suggested timelines for professional-development work, and protocols that can be given to teachers or placed in files that can be easily translated into systemwide databases and handouts with suggested competence-building techniques. Few, if any, observation systems currently provide these types of resources.

Different school systems have different resources available to devote to classroom observation. Some schools have personnel available to spend full days in classrooms in order to obtain data on important aspects of classroom functioning. Other school systems have less time available on a per classroom basis. In selecting an observational assessment instrument, it is vitally important that the instrument is used in practice in the same standardized ways it was used in development in order to obtain results with the expected levels of reliability and validity. Some instruments have been tested and validated using longer periods of observation than others. For that reason users may wish to generate a realistic approximation of how they will allocate observation time before selecting an assessment tool to ensure that the instrument selected can be used reliably and with validity within the parameters of that time budget.

In selecting an observational assessment instrument, it is vitally important that the instrument is used in practice in the same standardized ways it was used in development in order to obtain results with the expected levels of reliability and validity.

Different systems of observation require different time commitments. The amount of time that the observer will have available to them can be an important practical consideration when selecting an observational system. Keep in mind that in general, the more ratings one is able to obtain and aggregate, the more stable an estimate of typical teacher practices one will have. Most observational systems reporting sufficient levels of reliability and validity require a substantial amount of time for observation (at least one hour). If these types of validated tools are used, then ways must be found to accommodate these time demands. There is clearly a need for validated observational tools that can be completed quicker, particularly to accommodate the more typical observational strategies used by principals (which may be 5- or 10-minute walkthroughs), but none are currently available that meet the criteria reviewed above.

With regard to time of day, there is some evidence that, at least in elementary schools, observations completed during the first 30 minutes of the school day may yield lower ratings on some aspects of teaching, such as instructional practices, than observations conducted during the rest of the day. This isn't surprising given that this initial period of the day is typically used to complete management activities such as taking attendance and listening to school announcements. There is also some evidence that the quality of some social aspects of the classroom environment, such as classroom climate, may decrease over the course of the school day, which may reflect teacher and student fatigue. Other aspects of teaching practice, like instruction, seem to be more consistent after the first 30 minutes of the school day. Users of classroom observations may wish to consider these factors when deciding when to observe. There may be good reasons to observe during the beginning of the school day, however, if scores on observations are going to be used to compare teachers, a good policy may be to standardize the observational protocol to either include or not include these first 30 minutes.

With regard to time of year, findings from observations throughout the school year indicate that by and large there is consistency in teachers' behaviors over time, but there are indications that in general scores are somewhat lower at the very beginning of the year, around the winter holidays, and again at the end of the school year. For these reasons it is advisable to avoid the first and last months of the school year and days leading up to the winter holidays if the objective is to obtain scores that accurately represent typical practice.

Summary: Choosing and using observational protocols

While it may not always be possible to find tools that meet all the criteria we've outlined, it is nonetheless important that users evaluate potential observation systems with these criteria in mind and consider ways to address areas of concern. (Consider pilot testing and data gathering if an instrument hasn't been evaluated as a predictor of your specific outcomes of interest).

Above all, users must understand the types of inferences that are appropriate based on the data collected. Observational data can support inferences related to identifying teacher classroom behaviors that matter for students, describing typical practices in classrooms, determining how a given classroom or teacher compares with a national or district average, predicting what is the teacher's likely contribution to children's learning, and determining the extent to which teachers' practices improve in response to professional development. In order to draw any conclusions from observational data, however, the instruments must be subjected to extensive testing and evaluation. Users must be cautious to not overstep the appropriate use of observational instruments.

There is currently very little data to indicate the appropriateness of cut-off scores that would separate "sufficient" from "insufficient" levels of teaching skill on any of the reviewed instruments. Likewise, there are no published norms to guide expected levels of change in response to a given intervention strategy over a given period of time. For these reasons we must be extremely cautious in using observational data to determine whether teachers pass or fail in their provision of quality teaching or whether their progress in response to intervention is sufficient or lacking. In the future, with additional research, these types of inferences are likely to be more tenable. For the time being, however, the most appropriate use of observational data is to provide a sense of individual or programmatic areas of strength and areas of challenge, to guide individualized professional development or other support, and to determine if that support is working to move teachers "up" in their ability to provide quality teaching.

Using observation data to systematically improve the quality of classroom practice

Certainly the goal is to use observational methodology and the data acquired from observations to help teachers meet the challenges they face and in so doing improve the quality of their classroom practice. Creating a highly effective professional-development system is a sizable task that requires orienting efforts toward ongoing, individualized support for teachers to produce specific practices that impact students' growth and development.[33] This is a significant shift from the current standard—a workshop-based, one-size-fits-all approach.

Professional development is most effective if it is constructed around helping teachers make improvements in areas that really matter for students, when those areas targeted for observation and improvement are clearly defined, and when all participants agree that the targets of the observation are valid goals to work toward.

Selecting an observational tool that has demonstrated associations between observation-based scores and high-priority aspects of student development is helpful in getting all participants on the same page on what is being observed and why. The behaviors being observed can be directly translated into goals for practice. The language used by the tool provides members of an organization with a shared vocabulary and an underlying understanding of program goals along with facilitating clear communication and collaboration.

Enhancing the teacher-observer relationship

Mr. Jones, a teacher, feels slightly anxious as he anticipates the arrival of Dr. Taylor, his assigned staff-development professional. He has had contact with Taylor only once before, at the first of his two yearly observational assessments. Taylor called in advance to arrange a time to observe, but called this morning to say he would be delayed and the he would try to make it in the afternoon. Jones understands that delays can be unavoidable but he had prepared his whole morning so that Taylor would be able to observe him testing out new strategies that he wants specific feedback about.

When Taylor finally arrives he is friendly and courteous, but seems rushed and departs after only a brief observation. He leaves a copy of his evaluation for Jones to read with a note thanking Jones for his time. The evaluation, however, fails to touch on the areas of most concern to Jones and doesn't provide the direction he was seeking because there was no lead-in conversation between Jones and Taylor. Jones wishes that he had had the opportunity to share his thoughts with Taylor rather than being "tested" by a system that was not individualized to meet his specific professional needs. What's more the evaluation provides no concrete suggestions for fine-tuning Jones's practice or links to the specific behaviors engaged in by Jones that would have resulted in determinations of "needs attention," "meets expectations," or "does not meet expectations." Overall, Jones does not find the results of the evaluation particularly useful.

For another teacher, Mr. Lee, the experience of being observed was very different. At the start-the-school-year in-service meetings, all teachers received an orientation to the observational system that the school would be using to evaluate teachers. This orientation allowed teachers to get a sense about what kinds of teaching behaviors were important to incorporate into their practice and how they could expect those practices to impact students. Teachers were then paired with coaches who also gave brief overviews that included outlines of the professional-development system and how it would work. The coaches then met with individual teachers one-on-one to hear about their personal goals for the year as they related to the practices that would be assessed in the classroom observations. Coaches tried to visit classrooms on request as well as on a monthly basis. The classroom observations and feedback were focused on the specific goals that teachers had set for themselves at the start of the year or on new goals that teachers and coaches had set in response to observational findings or teachers' requests for assistance.

Lee was observed on several occasions by his coach Ms. Brown who gave him feedback about specific behaviors in written form. Each observation was followed up with a face-to-face meeting or phone calls shortly afterwards to review Brown's feedback, get Lee's perspective, and brainstorm specific ideas for making positive changes. Each meeting ended with Lee and Brown deciding together on the areas where Lee might best focus his efforts prior to the next observation. During that next observation the areas previously identified would be honed in on. Unlike Jones's experience, Lee feels that his coach/observer is a great resource and the good working partnership allows Lee to reflect on his work in a more focused and productive way.

Observational data only contributes to professional-development efforts if it is shared effectively with teachers. Giving teachers feedback about the results of observations and helping teachers reflect on this feedback in productive ways provides the bridge between knowledge about what matters for students and changes in teachers' actual practice. Both the content and style with which feedback is communicated are important areas to consider. Our recommendation, stemming from successful observationally based professional-development initiatives, is that feedback is most effective when it is: focused on increasing a teacher's own powers of observation, promotes reflection and self-evaluation skills, promotes intentionality around behaviors and patterns of interaction with students, helps teachers see the impact of their behaviors more clearly, and assists teachers in improving their implementation of lessons and activities. Doing this means providing feedback that is specific and behavioral in nature and balances attention to a teacher's positives and strengths with constructive challenges.

Focusing observations to improve outcomes

Student teacher Ms. McIntyre was formally observed by her lead teacher, Dr. Douglas, on three occasions. Following the first observation, the two met to discuss Douglas's feedback. In her observation Douglas used a system that included five broad areas of practice, each of which including 7 to 10 subcategories.

Douglas diligently went through McIntyre's level of performance in 43 areas. Because there are so many areas, Douglas felt that she only had time to touch on the level of proficiency that McIntyre demonstrated in each area without going into detail or giving many examples of specific behaviors observed. Both Douglas and McIntyre were dissatisfied with the process. Additionally, McIntyre was unsure how to improve in areas where she lacks confidence.

During the second observation Douglas decided to focus her feedback only on an area of exceptional strength for McIntyre and on an area with which she struggles. Although all 43 areas of practice were observed, the feedback was much more directed. In the follow-up conversation of this observation Douglas was able to give specific examples of the kinds of teacher and student behaviors she observed. She shared with McIntyre exactly how specific responses to students' comments increased engagement as well as how missing early signs of student disengagement resulted in time being taken away from instruction and instead directed to behavior. While this observational experience felt more helpful to both parties the issue of missed early signals of disengagement failed to resonate with McIntyre, precisely because she had missed them.

To remedy this shortcoming, for the next observation Douglas and McIntyre agreed to videotape the lesson so that they can review the tape together and see the exact same behavioral exchanges. Taking this approach allowed McIntyre to see exactly where she needed to shift her attention and pinpointed changes she could make in her physical presence in the classroom (moving around versus always standing at the front of the room), in the frequency with which she scanned the room, and in how she responded when she noticed a student who appeared bored. Again, Douglas still rated all 43 areas of practice if needed, but this kind of focused feedback supported by the use of video footage was much more helpful to McIntyre than simply reviewing large numbers of scores.

Certainly, making a single observation and providing feedback is a useful start, but to be effective the observation-feedback cycle needs to be repeated multiple times over the course of a school year. The aim should be to build on the lessons of the first observation and carrying those lessons forward into subsequent observations so that initial feedback is specifically addressed in follow-up observations. Just as teachers are encouraged to do formative assessments with their students in order to help them learn, this type of formative assessment of teachers' practices can help them recognize and improve their instruction. Similar to formative assessments of student learning, teachers and support personnel can use data from observations to guide planning for making changes and to guide the selection of behaviors that will be the focus of follow-up observations. This process of feeding data back into the system maximizes the effectiveness of efforts toward improvements in the teaching practice. Charting progress, being able to document systematic progress towards goals (or lack thereof), and recording agreed upon strategies for making changes all help make observational data a highly effective tool for providing support for professional development.

Recommendations and lessons derived from observation in early childhood education

The experience with scaling-up observational assessments in early childhood education demonstrates that standardized observational approaches used to measure teacher performance represent a credible complement to the current focus on teacher credentials and degrees on the one hand and the value-added metrics of student performance on the other. Furthermore, observational approaches link more directly to professional-development systems for producing effective teaching and as such represent an alternative to credentials or degrees that may have greater long-term benefits for building capacity and quality. Below are a set of key lessons learned from work in early childhood education that may have utility for K-12 educators as they launch into the use of observational measures of teacher performance as well as for policymakers and district leaders who advocate such uses.

- Any measure must provide information in the form of metrics that discriminate among those being assessed if such measures are going to be useful in any form of decision making. Observation is no exception, thus observation should be a form of measurement and assessment consisting of codes and benchmarks applied rigorously, just as they are in assessments of student performance.

- Observations used in systems of decision making and performance improvement at any level of scale must adhere to standardized procedures. There are three components of standardization that are key elements for evaluating any observation instrument and its implementation: training protocol, parameters around observation, and scoring directions.

- The technical properties of observational protocols and scoring systems are fundamental for their use. Reliability is one of these properties and pertains to the level of random error or bias in the scores obtained. It is critical that users select tools that have documented reliability for use across observers, teachers, time, and situations when metrics obtained from these tools will be used to draw conclusions about teacher performance. Effective training programs for observers help ensure raters are consistent with one another as they make ratings. Similarly, including periodic "drift" testing at predetermined intervals will help to improve the degree to which raters remain consistent with scoring protocols and with each other.

- Any observation of teacher performance must show empirical relations with student learning and development if the use of observation is expected to drive improvement in student outcomes. Selecting an observation system that includes validity information cannot be overstated.

- Pragmatically, observation takes time and different systems of observation require different time commitments. The amount of observer time available can be an important practical consideration when selecting an observational system. In general, the more ratings a school or district is able to obtain and aggregate the more stable an estimate of typical teacher practices will result.

- Observations can identify teacher classroom behaviors that matter for students, describe typical practices, show how a given classroom or teacher compares with a national or district average, forecast the likely contribution of a teacher to children's learning, or document improvement in teachers' practices in response to professional development. Users, however, must be cautious to not overstep the appropriate use of observational instruments in their enthusiasm to apply them in any and all circumstances.

- Observations can be used in both accountability and program-improvement applications. Importantly, policy and program investments can over time change the typical distribution of scores as teachers, classrooms, and programs improve. As a consequence it can be necessary to periodically "raise the bar" on performance standards or cutoff scores.

- Feedback to teachers is most effective when it is individualized, highly specific, and focused on increasing teacher observation skills, promoting self-evaluation, and helping teachers see and understand the impact of their behaviors more clearly.

The evidence from years of classroom observation in early childhood education suggests that a teacher's performance in a classroom, in terms of actual behavioral interactions with students, can be assessed observationally in scaled-up applications using standardized protocols; can be analyzed systematically with regard to various sources of error; and in turn can be shown to be valid for predicting student learning gains as a function of specific and aligned supports provided to teachers. Exposure to such supports is predictive of greater student-learning gains.[34] The widespread introduction of observations into K-12 represents a tremendous opportunity and a massive challenge to a system not accustomed to doing this type of evaluation well.

K-12 educators would do well to learn from the lessons and experience accrued by their counterparts in the early childhood sector. At a time of considerable urgency and demand for school improvements the good news is there is no need to reinvent the wheel. In fact more explicit acknowledgement of the expertise already present in early childhood education might actually help K-12 educators proceed cautiously and thoughtfully, yet move with deliberate speed as they travel along this promising path of school improvement.

About the author

Robert Pianta is the dean of the Curry School of Education, the Novartis U.S. Foundation professor of education, and a professor of psychology at the University of Virginia, where he also directs the University of Virginia Center for Advanced Study of Teaching and Learning. His research and policy interests focus on the measurement and production of effective teaching in classrooms from preschool to high school. Pianta has published more than 300 scholarly papers and is lead author on several influential books related to early childhood and elementary education. He is the senior author and developer of the Classroom Assessment Scoring System, a method for assessing teacher/classroom quality being used in many district-, state- and national-level applications.

Pianta's assessments of teacher effectiveness are the national standard for Head Start classrooms and are included in the Gates Foundation's "Measures of Effective Teaching" study. Pianta is principal investigator and director of the Institute of Education Sciences' National Center for Research on Early Childhood Education; is principal investigator of MyTeachingPartner, a NICHD-funded clinical trial evaluation of Web-based coaching and video training for PreK teachers; and co-principal investigator on MyTeachingPartner-Secondary, an evaluation of this approach in middle school and high school classrooms.

Pianta is co-director of the University of Virginia's interdisciplinary pre- and post-doctoral training programs in education sciences and past editor of the *Journal of School Psychology*. He consults with numerous foundations as well as state and federal initiatives.

Acknowledgements

The author wishes to acknowledge the generous support of the William T. Grant Foundation, which provided resources for the case examples and the descriptions threaded throughout this report. In addition, the author wishes to thank his colleagues Bridget Hamre and Jason Downer, who have for many years been part of the efforts to use observations to improve teaching. Also deserving of appreciation and recognition are numerous longstanding colleagues at the Center for Advanced Study of Teaching and Learning at the University of Virginia, including Marcia Kraft-Sayre, Kathy Neesen, and Jeanne Stovall and the many collaborators at other institutions who have been part of our work.

Endnotes

1 Bill & Melinda Gates Foundation, "Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project" (2010); National Council on Teacher Quality, "Increasing the Odds: How Good Policies Can Yield Better Teachers" (2005).

2 Bill & Melinda Gates Foundation, "Learning about Teaching"; Robert C. Pianta, "Standardized Observation and Professional Development: A Focus on Individualized Implementation and Practices." In M. Zaslow and I. Martinez-Beck, eds., *Critical Issues in Early Childhood Professional Development* (Baltimore: Brookes, 2005).

3 Bill & Melinda Gates Foundation, "Learning about Teaching."

4 Robert C. Pianta and others, "The Effects of Preschool Education: What We Know, How Public Policy Is Or Is Not Aligned with the Evidence Base, and What We Need to Know," *Psychological Science in the Public Interest* 10 (2009): 49–88.

5 Pianta, "Standardized Observation and Professional Development."

6 T. Harms and R. M. Clifford, *Early Childhood Environment Ratings Scale* (New York: Teachers College Press, 1980); T. Harms, R. M. Clifford, and D. Cryer, *The Early Childhood Environment Ratings Scale: Revised Edition* (New York: Teachers College Press, 1998).

7 Robert C. Pianta, Karen La Paro, and Bridget K. Hamre, *Classroom Assessment Scoring System (CLASS)* (Baltimore: Paul H. Brookes, 2008).

8 Pianta and others, "The Effects of Preschool Education," 49–88; A. W. Mitchell, "Models for Financing State-Supported Prekindergarten Programs." In R. Pianta and C. Howes, eds., *The Promise of Pre-K* (Baltimore: Brookes, 2009).

9 Harms and Clifford, *Early Childhood Environment Ratings Scale*; Harms, Clifford, and Cryer, *The Early Childhood Environment Ratings Scale: Revised Edition*.

10 J. Ludwig and D. L. Miller, "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design," *Quarterly Journal of Economics* 122 (2007): 159–208.

11 Pianta and others, "The Effects of Preschool Education"; Ludwig and Miller, "Does Head Start Improve Children's Life Chances?"; Administration on Children and Families, *FACES Findings: New Research on Head Start Outcomes and Program Quality* (Department of Health and Human Services, 2006); Administration on Children and Families, *Head Start Impact Study: Final Report* (Department of Health and Human Services, 2010).

12 Mitchell, "Models for Financing State-Supported Prekindergarten Programs."

13 Pianta and others, "The Effects of Preschool Education."

14 National Association for Regulatory Administration, *The 2007 Child Care Licensing Study* (Lexington: Author and National Child Care Information and Technical Assistance Center, 2009); American Public Health Association and the American Academy of Pediatrics, "Caring for our Children: National Health and Safety Performance Standards: Standards for Out-of-Home Child Care Programs" (1992); J. Layzer and C. Price, *Closing the Gap in the School Readiness of Low-Income Children* (Department of Health and Human Services, 2008).

15 Administration on Children and Families, *FACES Findings*.

16 Pianta and others, "The Effects of Preschool Education"; American Public Health Association and the American Academy of Pediatrics, "Caring for our Children"; Layzer and Price, *Closing the Gap in the School Readiness of Low-Income Children*.

17 Pianta, Paro, and Hamre, *Classroom Assessment Scoring System (CLASS)*.

18 Pianta and others, "The Effects of Preschool Education"; Andrew Mashburn and others, "Measures of Classroom Quality in Prekindergarten and Children's Development of Academic, Language, and Social Skills," *Child Development* 79 (2008): 732–749.

19 National Institute of Child Health and Human Development Early Child Care Research Network, "Child-Care Structure > Process > Outcome: Direct and Indirect Effects of Child-Care Quality on Young Children's Development," *Psychological Science* 13 (2002): 199–206; Robert C. Pianta and others, "A Day in Third Grade: A Large-Scale Study of Classroom Quality and Teacher and Student Behavior," *The Elementary School Journal* 105 (2005): 305–323.

20 Mashburn and others, "Measures of Classroom Quality in Prekindergarten and Children's Development of Academic, Language, and Social Skills"; R. Clifford and others, "What is Pre-Kindergarten? Trends in the Development of a Public System of Pre-Kindergarten Services," *Applied Developmental Science* 9 (2005): 126–143.

21 Bill & Melinda Gates Foundation, "Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project"; Mashburn and others, "Measures of Classroom Quality in Prekindergarten and Children's Development of Academic, Language, and Social Skills"; Robert C. Pianta and others, "Features of Pre-Kindergarten Programs, Classrooms, and Teachers: Do They Predict Observed Classroom Quality and Child-Teacher Interactions?", *Applied Developmental Science* 9 (2005): 144–159.

22 Robert C. Pianta and others, "Effects of Web-Mediated Professional Development Resources on Teacher-Child Interactions in Pre-Kindergarten Classrooms," *Early Childhood Research Quarterly* 23 (4) (2008): 431–451; Joseph P. Allen and others, "An Interaction-Based Approach to Enhancing Secondary School Instruction and Student Achievement," *Science* 333 (6045) (2011): 1034–1037.

23 Bill & Melinda Gates Foundation, "Learning about Teaching"; Pianta and others, "Effects of Web-Mediated Professional Development Resources on Teacher-Child Interactions in Pre-Kindergarten Classrooms"; Allen and others, "An Interaction-Based Approach to Enhancing Secondary School Instruction and Student Achievement."

24 Bill & Melinda Gates Foundation, "Learning about Teaching"; Robert C. Pianta and Bridget K. Hamre, "Conceptualization, Measurement, and Improvement of Classroom Processes: Standardized Observation Can Leverage Capacity," *Educational Researcher* 38 (2) (2009): 109–119.

25 Bill & Melinda Gates Foundation, "Learning about Teaching"; National Association for the Education of Young Children, "Position Statements of NAEYC" (2005).

26 Bill & Melinda Gates Foundation, "Learning about Teaching"; Harms and Clifford, *Early Childhood Environment Ratings Scale*; Harms, Clifford, and Cryer, *The Early Childhood Environment Ratings Scale: Revised Edition*; Pianta, Paro, and Hamre, *Classroom Assessment Scoring System (CLASS)*; Pianta and Hamre, "Conceptualization, Measurement, and Improvement of Classroom Processes: Standardized Observation Can Leverage Capacity."

27 Bill & Melinda Gates Foundation, "Learning about Teaching"; Pianta and Hamre, "Conceptualization, Measurement, and Improvement of Classroom Processes: Standardized Observation Can Leverage Capacity."

28 Ibid.

29 Bill & Melinda Gates Foundation, "Learning about Teaching."

30 Bill & Melinda Gates Foundation, "Learning about Teaching"; Harms and Clifford, *Early Childhood Environment Ratings Scale*; Harms, Clifford, and Cryer, *The Early Childhood Environment Ratings Scale: Revised Edition*; Pianta, Paro, and Hamre, *Classroom Assessment Scoring System (CLASS)*; Pianta and Hamre, "Conceptualization, Measurement, and Improvement of Classroom Processes."

31 Bill & Melinda Gates Foundation, "Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project."; Allen and others, "An Interaction-Based Approach to Enhancing Secondary School Instruction and Student Achievement."; Pianta and Hamre, "Conceptualization, Measurement, and Improvement of Classroom Processes.".

32 Pianta and others, "The Effects of Preschool Education"; Pianta and others, "Effects of Web-Mediated Professional Development Resources on Teacher-Child Interactions in Pre-Kindergarten Classrooms"; Allen and others, "An Interaction-Based Approach to Enhancing Secondary School Instruction and Student Achievement."

33 Pianta and others, "The Effects of Preschool Education"; A. W. Mitchell, "Models for Financing State-Supported Prekindergarten Programs"; Harms and Clifford, *Early Childhood Environment Ratings Scale*; Harms, Clifford, and Cryer, *The Early Childhood Environment Ratings Scale: Revised Edition*; Pianta, Paro, and Hamre, *Classroom Assessment Scoring System (CLASS)*; Pianta and others, "Effects of Web-Mediated Professional Development Resources on Teacher-Child Interactions in Pre-Kindergarten Classrooms"; Allen and others, "An Interaction-Based Approach to Enhancing Secondary School Instruction and Student Achievement."

The Center for American Progress is a nonpartisan research and educational institute dedicated to promoting a strong, just, and free America that ensures opportunity for all. We believe that Americans are bound together by a common commitment to these values and we aspire to ensure that our national policies reflect these values. We work to find progressive and pragmatic solutions to significant domestic and international problems and develop policy proposals that foster a government that is "of the people, by the people, and for the people."

Center for American Progress

1333 H STREET, NW, 10TH FLOOR, WASHINGTON, DC 20005 • TEL: 202-682-1611 • FAX: 202-682-1867 • WWW.AMERICANPROGRESS.ORG