How to Select The Right classroom OBSERVATION Tool

This booklet outlines key questions that can guide observational tool selection. It is intended to provide guiding questions that will help users organize their thinking about what they want from an observation tool and help them to find instruments well aligned with their strategic goals.

Part 3

of a 5 Part Series:

A Practitioner's Guide to Conducting Classroom Observations: What the Research Tells Us About Choosing and Using Observational Systems to Assess and Improve Teacher Effectiveness

CASTL
CENTER FOR ADVANCED STUDY OF TEACHING AND LEARNING

Megan W. Stuhlman, Bridget K. Hamre, Jason T. Downer, & Robert C. Pianta, University of Virginia

This work was supported by a grant from the WT Grant Foundation.

Choosing the Right Observational Tool: Factors to Consider

There are multiple published and unpublished classroom observation systems available for use, and deciding among them is the first step in putting an observational system to work in your organization.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

When reviewing such tools, the following questions can be used to guide the decision-making processes regarding which observation system is best suited to the needs of a particular organization.

Tier 1: High Priority Questions

- Has this tool been shown to produce reliable scores across observers and over time?

- Are the outputs (scores) from this observation protocol proven to relate to outcomes of interest in our population (i.e., growth in students' academic skills, students' prosocial behaviors, teacher retention, students' reports of feelings of belonging, etc.)? In other words, is the instrument valid for our intended purpose?

- What questions about classrooms does my organization want answered? Is the scope of this tool aligned with the questions about classrooms and teachers' practices that we want to address?

- Are the observation and scoring protocols standardized and clear?

Tier 2: Additional Considerations

- Does the system include complementary sources of information (such as student surveys, etc.) that could be used to obtain a more complete portrait of the classroom?

- Does the observation include guidelines and support for using findings for professional development purposes?

- Is the time required for observation feasible for your organization?

Each of these questions is reviewed in more detail below.

Does the observation include reliability information?

Instrument reliability is a key consideration in selecting an observational assessment tool. Instrument reliability means that whatever qualities a given tool is measuring, it should measure those qualities consistently. In observational assessments of classrooms, a tool that produces reliable scores will output the same score regardless of variation in the classroom that is outside of the scope of the tool and regardless of who is making the ratings.

For example, 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.

No observation of teaching practices will produce perfectly reliable scores. We know that despite high levels of training, observers will sometimes make different judgments. We also know that certain classroom activities may influence scores on observational tools.The goal is to choose an observational tool that can produce relatively high-reliability scores and to be aware of potential biases.

There are several aspects of reliability. Perhaps the two most relevant when considering classroom observation systems are stability over time and consistency across observers. With regard 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.

Key Concept –Reliability

Look for instruments that provide scores that are:

- Consistent over time **unless** change is expected.
- Consistent across observers.

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 feedback and support around these behaviors is much more likely to resonate with teachers and to function as useful levers for helping them change their practice. It is advantageous for observational tools to 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, etc).

A notable exception around the criteria of stability over time as a marker for reliability 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 an organization's curriculum is changing or new program-wide goals are being implemented, a lack of stability in observations of teacher behaviors may well represent true change 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 at scale, training protocols and provision of scoring directions must be clear and extensive enough to produce an acceptable level of 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 unknown.

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 for which there is documented consensus among trained raters on whether or not or to what extent teachers are engaging in the behaviors under consideration.

Does the tool provide information on validity?

Validity represents the degree to which the ratings produced by the observation system are associated with the student or teacher outcomes about which the observation is designed to provide information. 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 to show 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.

Key Concept –Validity

Look for instruments that provide scores with proven links to 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 they are gathering is relevant to the outcomes 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. We must know that our assessment tools are directly and meaningfully related to our outcomes of interest before we begin using them either in professional development or accountability frameworks.

A system may well be valid for one set of outcomes but not for another, so clarity around outcomes of interest is important. For example, an observation system may include validity data regarding the prediction of students' academic achievement during that school year, but it may demonstrate no relation to student drop-out 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 instrument may be well-suited for that purpose. However, if the objective is to determine whether teachers are enacting behaviors that will prevent drop-out, a different observation with documented links to drop-out rates may be preferable.

If a user has a particular observation tool that is well aligned with the questions they want answered about classroom practice and meets the criteria summarized previously, 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. For example, a district or organization 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 more 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, teacher educators and school personnel can feel confident establishing observationally-based standards and mechanisms for meeting those standards, and educational systems, teachers, and students will all benefit.

What questions about classrooms do I want answered? Do the scope and design of the instrument lend themselves to addressing these questions?

Scope of Observations. Different instruments provide users with different types of information about classrooms. Some are inclusive of multiple varied aspects of teaching practice, providing data on layers of setting quality including the physical environment, the types of activities observed in the classroom, and the teacher's execution of professional responsibilities such as record keeping and communicating with families.

Others adopt a highly focused approach, such as exclusively attending to a highly detailed and 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 not including information that would require knowledge outside of what is obtained during specified observation windows (i.e., not including how the teacher communicates with parents, makes lesson plans, etc.).

Users may wish to begin the selection process by defining the goals that their organization has in using an observation tool. After having defined the desired outcome, users can select a measurement tool that is well aligned with their objectives.

Age Range Covered. In addition to ensuring a match between the scope of what is assessed by the instrument and system goals, users are also advised to attend to the age range that the instrument was designed for and the grade levels from which data on the psychometric properties of the instrument have been obtained. For example, if your goal is to assess fourth-grade classrooms, 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.

Global Versus Content Specific. 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 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 your organization has a particular interest in a certain content area, you may wish to supplement a protocol for observing generalized supports with one that includes specific interactive practices relevant to your content area of focus.

CASE STUDY # 1:
Choosing an Observation Tool for a Specific Curricula

The 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 now wants to evaluate the extent to which teachers using this curriculum are incorporating high-quality strategies for teaching mathematics in comparison with the extent to which teachers in a control group of schools are incorporating such strategies in teaching mathematics in order to help them decide whether this curriculum may be a good choice for district-wide use.

This school district may wish to use an observation protocol 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.

CASE STUDY # 2:
Choosing an Generalized Observational Tool

The Lakeview school district wishes to conduct an observational assessment of all teachers in order to gain a better understanding of system-wide areas of strength and challenge so that they can plan for in-service programming and create individualized professional development plans for teachers. Observers will conduct multiple observations per day, so these observations will occur at different times of day and during different activities for different teachers.

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

CASE STUDY # 3:

Choosing an Observational Tool for Merit Pay and Tenure

Franklin County school district wants to outline a structure for merit pay and tenure decisions that includes quality of observed teaching behaviors as one of their components. Therefore, the county decides to select an assessment instrument that has shown a relationship to student outcomes at different levels of quality. In other words, one with research support demonstrating that incremental gains in the quality of the measured teaching practices result in incremental gains in student performance.

They then stipulate two options for sufficient practice in this component: 1) teachers demonstrate high-quality teaching practices in initial and follow-up assessments, or 2) teachers demonstrate improvement over time in quality of teaching practices/positive response to professional development support as indicated by increasing scores over time.

Global Rating Methodology Versus Frequency Counts of Behaviors. An additional consideration that falls in this scope category concerns the degree to which observational systems capture information on the frequencies of certain teacher behaviors or on more holistically defined patterns of behavior. Measures using *time-sampling methodology* ask users to count the number of specific types of behaviors observed. *Global rating methodology guides* users to watch for patterns of behavior and make summative judgments about the presence or absence of these behaviors.

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, and the extent to which classroom interactions contain a high degree of negativity, both between teachers and students and among peers.

There are advantages and disadvantages to each type of system. An advantage to global ratings is that they assess higher-order organizations of behaviors in ways that may be more meaningful than looking at the discrete behaviors in isolation. For example, teachers' positive emotions and smiling can have different meanings and may be interpreted differently depending on the ways in which students in the classroom respond. In some classrooms teachers are exceptionally cheerful, but their emotions appear very disconnected from those of the students. In other classrooms teachers are more subdued in their expressed positive emotions but there is a clear match between this level of emotional expression and that of the students.

Key Concept –Observational Methods
Time-Sampling Methodology/Frequency Counts: most adept at highlighting differences within a specific teacher's practices during different specific teaching activities.

Global Rating Methodology: most adept at highlighting stable teacher characteristics and at providing information that differentiates between teachers.

A measure that simply counted the number of times a teacher smiled at students would miss these more nuanced interpretations. However, an instrument characterized by time-sampling methods, with a focus on frequencies of specific behaviors, may lend itself well to easy alignment with the evaluation of certain interventions. For example, if a goal is to increase the numbers of times that teachers provide students with specific and focused feedback rather than giving no feedback or simply saying "yes" or "no," an instrument using time-sampling methods could provide very concrete data on the extent to which an intervention impacted this specific behavior by counting the frequencies of specific and focused feedback 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 specifically evaluated using time-sampling methods as well.

One other difference between these two approaches concerns the degree to which they are subject to observer effects. There tend to be more significant observer effects using global ratings 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 requires much less inference than does making a holistic judgment about the degree to which a teacher fosters a positive classroom climate.This point emphasizes the need for adequate training and strategies for maintaining reliability among classroom observers, issues considered in greater detail in the next sections.

Another factor to consider is how much of the variance in these ratings can be attributed to stable characteristics of the classroom versus factors that change over time as a result of subject matter, number of students, time of day, etc. Evidence suggests that time-sampled codes show little classroom-level variance, in contrast to global ratings, in which the bulk of the variance was at the classroom level. This indicates that the time-sampled codes are not as sensitive to differences between teachers and classrooms as are the global ratings. This is an important consideration for users interested in obtaining information about different teachers' individualized strengths and areas of challenge.

Is the instrument standardized in terms of administration procedures? Does it offer clear directions for conducting observations and assigning scores?

Once you have clarified your purpose and goals in conducting classroom observations, 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. This is an essential component of a useful observation system: 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, and thus hampers system-wide applicability.

There are three main components of standardization that users may consider evaluating in an observation instrument:

1. training protocol;
2. observation protocol;
3. scoring directions

Training Protocol. With regard to the training protocol, 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 (i.e., do all observers need to pass a reliability test, observe in a certain number of classrooms, be consistent with colleagues at a certain level)?

Observation Protocol. 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?), direction around time of day or specific activities to observe, as well as whether observations are announced or unannounced, and other related issues.

Scoring Directions. With regard to scoring, users are advised to look for clear guidelines. 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 (i.e., high, moderate, 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?

CASE STUDY # 4:
Importance of Observational Protocols

A teacher preparation program is looking for a way to assess students' performances at the beginning and end of their student teaching work, during which time they are also taking a course on effective teaching practice. They 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 the broader goals of the program, and therefore would be good targets for assessment. However, the system does not include training or observational protocols or explicit directions for scoring. As a consequence, it is used quite differently by two faculty members.

When Professor Jones makes observations, he has arranged 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. He ends the observation as the teacher ends the lesson. He takes detailed notes about the teachers' practice along the six dimensions. When scoring, he reasons that if he sees teachers engaging in the behaviors under consideration several times, they should get "full credit," or a 10, on the scale. Professor Allen 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 Jones' scores could be systematically higher than Professor Allen's.

We can see from this example that even with well defined and theoretically sound scales, a clear observation and scoring protocol that all observers follow is extremely important in terms of obtaining scores that are consistent across observers. In this example, note that significantly different scores are likely to result from Professor Jones' observations and Professor Allen'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 example, if Professor Jones used his interpretation of the protocol to conduct initial start-of-student-teaching observations and Professor Allen used her interpretation of protocol to conduct the end-of-student-teaching observations, any true gains in teaching practice could be obscured, and 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.

The four preceding factors represent key areas to consider when selecting an observation tool. Above and beyond these core factors, other potential considerations include:

Does the system include complementary sources of information?

Obtaining information about classrooms from multiple sources and from different perspectives (e.g., the teachers' own perspective, students' perspectives, perspective of someone generally familiar with the classroom on a routine basis) can provide a more comprehensive picture of the classroom environment. This can also be helpful in terms of providing constructive feedback – one could seek out coherent patterns in responses across observers/raters.

For example, having a teacher engage in a self-study or self-assessment in conjunction with structured observations made by neutral observers may be 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.

Key Concept –Standardization Procedures

Observations should be standardized around:

- Training protocol
- Observation protocol
- Scoring directions.

Does the observation include guidelines and support for using findings for professional development purposes?

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), choosing 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, protocols that can be given to teachers, placed in files, and be easily translated into system-wide databases and handouts with suggested competence-building techniques. Few observation systems provide these types of resources at this time.

Is the time demand for conducting the observation workable within my system?

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. Users may wish to generate a realistic approximation of how they will be able to allocate observation time before selecting an assessment tool. An instrument that can be used reliably and with validity within the parameters of that time budget can then be selected.

UNIVERSITY of VIRGINIA
CURRY SCHOOL OF EDUCATION

The University of Virginia Center for Advanced Study of Teaching and Learning (CASTL) focuses on the quality of teaching and students' learning. CASTL's aim is to improve educational outcomes through the empirical study of teaching, teacher quality, and classroom experience from preschool through high school, with particular emphasis on the challenges posed by poverty, social or cultural isolation, or lack of community resources.