

How To Use  
**classroom**  
OBSERVATION  
Most Effectively

This booklet includes a list of guiding principles for the successful use of observation tools, as well as logistical information regarding important ways to standardize observation protocols.



## Part 4

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





## Using Classroom Observational Systems Effectively

There are several ways that school systems can maximize the usefulness of observational assessments of classrooms. Most of these guidelines revolve around being aware of the needs of your individual system and selecting an observation tool that is well aligned with these needs, implementing standardized observation practices, and minimizing variation in environmental factors and rater effects that may impact observational data. Seven guiding principals are explored in greater detail below:

1. Be intentional about your choice of an observation system.
2. Know your observation tool and communicate this knowledge system-wide.
3. Know the appropriate uses of observational data.
4. Stick to the system.
5. Build in enough time, at the right times of day and times of year, to obtain stable, representative scores.
6. Minimize rater effects and other sources of error.
7. Document everything.

### Be intentional about your choice of an observation system.

Be aware of all of the options available to you as you begin the process of selecting an observation system for your organization and take the time to select an observation system that is closely aligned with your organization's goals and philosophy. Everyone who is influenced by the incorporation of observational methodology, particularly teachers, will appreciate congruence between stated program goals and the types of behaviors that are assessed in the chosen observation system. In a similar vein, choosing an observation system that has as much validity information as possible will help all participants to see the value of embracing the system and working to implement the behaviors that it assesses.

### Know your observation tool and communicate this knowledge system-wide.

Spending sufficient time getting to know the instrument that you have chosen is a crucial step. All protocols have areas of strength and weakness; having a thorough understanding of what the instrument you select does *and* does not accomplish and communicating this openly and clearly to those involved in its use is very important for participants as they become familiar with the new tool.

Observation at its best functions to join administrators, observers, teachers, and even students together around clearly articulated goals for practice and outcome. Keeping lines of communication open with accurate information is crucial to making the observation system you select a mechanism for bringing colleagues together and creating a clear, shared vision.

Along these lines, it is ideal if at least one person in your system is available to play the role of resident expert in the observation tool being used. It is very helpful if participants at all levels (teachers, evaluation personnel, mentors, directors, principals, etc.) have access to an on-site resource person who can be approached with questions and concerns that may arise as the instrument is being used. The resident expert can be the "point person" who coordinates use of the instrument, assists with standardization and fidelity of implementation, and perhaps is in contact with the developers of the instrument in order to obtain the most up-to-date information about the best ways to use the tool and interpret the results. Keep in mind that the developers or publishers of assessment systems may have additional resources that they can offer your organization, or may be able to provide specialized trainings tailored to your needs.

### Know the appropriate uses of observational data.

It is extremely important to understand the types of inferences that are appropriate to make once you have selected your observation tool and collected data. Observational data has the potential to expand our understanding of which aspects of teachers' classroom behavior matter for students, what typically happens in classrooms around the country, how what is happening in a given classroom compares with a national average, and how teachers respond to professional development or other supports designed to improve practice.

However, to draw any conclusions from observational data, the instruments we are using must be subjected to extensive testing and evaluation. The research community is just beginning to subject classroom assessment tools to that type of useful scrutiny. We must be cautious, therefore, as users, that we do not overstep the appropriate use of observational instruments in our enthusiasm to use them to their full, eventual potential.

For example, at this time, the evidence is clear that "higher" is better on reliable and validated measures of classroom quality. However, there is very little data to indicate the appropriateness of cut-off scores that would separate "sufficient" from "insufficient" levels of teaching skill on currently available instruments.



Likewise, at this time 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 get a sense of individual or programmatic areas of strength and areas of challenge, to guide individualized professional development or other support, and to check to see if this support is working to move teachers “up” in their ability to provide quality teaching.

### **Stick to the system.**

Once all participants know the system well, it is important to be very clear that the standard procedures outlined by the system should be followed without exception. If there are less clear procedural guidelines reviewed in the user’s guide, your organization as a whole may wish to determine appropriate internal guidelines and ensure that these are followed consistently by all observers.

Standardized and uniform practice is crucial to obtaining results or outputs that are useful for making comparisons across classrooms, creating system-wide professional development goals and support systems, and routing teachers on accurate, individualized pathways to receiving support in areas of difficulty.

Observers can be tempted to deviate from standardized practice when they feel that an observation period was not representative of a teacher’s typical practice, when they are concerned that a teacher will look “too bad” or “too good” or simply when they are short on time and are trying to fit in more observations during non-standard time periods or spending less than the required time to observe. Some of these situations can be addressed by emphasizing the importance of clear guidelines around such factors as time of day, length of observation, and the like, and explaining why this is necessary in order to have accurate, fair information across teachers.

It is also helpful to include space on the observation form or scoring sheet where observers can record information such as time of day and activity, as well as a space for them to note whether they felt that the results of the observation accurately captured their subjective impression of the quality of the classroom. Observers often appreciate being provided with a format for documenting any factors that they felt skewed the observation results if they have concerns. Providing this kind of outlet may reduce any inclination to deviate from standard procedures when observing or assigning scores.

It can also be helpful to explain that it is always better to stick to the observation protocol and possibly redo the observation at a later time than to assign scores or conduct observations in non-standard ways. Also, reviewing the validity information on the observation tool can be a useful reminder that, when completed according to standard protocol, these observations are in fact useful for predicting outcomes of interest. When completed in non-standard ways, they represent an unknown, which is not useful for the teacher, the students, or the larger system.

### **Build in enough time, at the right times of day and times of year, to obtain stable, representative scores.**


Different systems of observation require different time commitments. The amount of time that the observer will have available can be an important practical consideration when selecting an observational system. Keep in mind that in general, the more ratings you are able to obtain and aggregate, the more stable an estimate of typical teacher practices you will have. This may be particularly true for practices related to the provision of instructional support, which shows evidence of greater variability across a day than emotional support or classroom organization.

Most of the observational systems reporting sufficient levels of reliability and validity require a substantial amount of time for observation (at least one hour, often significantly more), or a significant number of observation cycles to be completed. If you want to obtain sufficiently reliable data using these types of validated tools, you will need to think through ways to accommodate these time demands.

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 of teaching quality 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. This may reflect teacher and student fatigue. Other aspects of teaching practice (e.g., instruction) seem to be more consistent after the first 30 minutes.

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 (e.g., to



observe the way a teacher handles management routines); 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, initial 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 at the very end of the school year. For these reasons, if your objective is to obtain scores that accurately represent typical practice, it is advisable to avoid the first and last months of school and days leading up to the winter holidays if possible.

## Minimize rater effects and other sources of error.

Rater effects may seem particularly relevant if there will be multiple people conducting observations within a given system. However, even if a single individual is conducting all observations within a school, and 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 take periodic "drift" tests to ensure reliability with the standards outlined by the developers of the measure (i.e., 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 and with one another. Other sources of error to consider include the extent to which ratings may represent factors specific to the day or time that was observed and not reflective of teachers' typical practices. Several steps that can be taken to minimize rater bias and other sources of error in measurement of typical practices are reviewed below.

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 as they make ratings.

Including periodic "drift" testing or calibration at predetermined intervals can offer a refresher in scoring procedures and help improve the degree to which raters remain consistent with scoring protocols and with each other.

### Key Concept – Reliability Across Raters

Rater effects can be minimized by:

- Training
- Testing
- Sticking to standardization
- Including frequent opportunities for raters to calibrate their scores
- Random assignment of observers to classrooms
- Systematic rotation of observers
- Increase numbers of observation cycles/days
- Two or more observers per classroom (not necessarily observing simultaneously)

With regard to scheduling observations/assigning raters to classrooms, rotating raters across teachers and including multiple raters for each teacher can help avoid systematic variance in scores. Visiting all classrooms on two or more days over the course of the year is good practice, and scheduling observations so that two different observers have the opportunity to rate each classroom on each of those observation days will produce more reliable estimates of typical teacher practices in those classrooms.

For example, if Vice Principal Smith and Curriculum Coordinator Jones share observation responsibilities, consider having each rater observe Classroom A and Classroom B on each of two observation days, alternating such that on day 1, Vice Principal Smith begins in Classroom A for two cycles of observation, then switches to Classroom B for the two subsequent observation cycles. Curriculum Coordinator Jones begins in classroom B on that day, switching after two cycles to Classroom A. They could repeat this procedure in reverse on the second scheduled observation day. Initial analysis of observation data collected in several different scheduling styles has suggested that this schedule (diagramed on page 5) may produce the highest levels of reliability using the fewest observation cycles.

Random assignment of observers to classrooms can also be useful in reducing systematic rater bias. Alternately, multiple raters can observe and rate classrooms simultaneously and their scores can be averaged, thus reducing the amount of bias introduced by any single observer who may be a systematically high or low rater.



## Assigning multiple observers to classrooms to maximize reliability of ratings

		Classroom A				Classroom B			
Rater		Cycle 1 8:30 a.m.	Cycle 2 9 a.m.	Cycle 3 9:30 a.m.	Cycle 4 10 a.m.	Cycle 1 8:30 a.m.	Cycle 2 9 a.m.	Cycle 3 9:30 a.m.	Cycle 4 10 a.m.
Day 1	VP Smith								
	CC Jones								
Day 2	VP Smith								
	CC Jones								

## Document everything!

As a general rule, we recommend systematically recording any factors that may have bearing on the results of the observation along with the assessment scores. These environmental factors external to the observational ratings themselves can be treated as data in the same way scores are treated as data.

Some examples of factors to record include: time of day, time of year, activity (math, science, social studies, managing classroom business), social setting in the classroom (numbers of adults and children, grouping), whether the typical members of the classroom are present, whether it is a typical day, and other environmental factors deemed relevant.

Ideally, these factors are kept as consistent as possible across observations. Given the demands of observing large numbers of classrooms, sometimes this is not feasible. If it is not possible to standardize these aspects of the observation, documenting

the content makes it possible to include this information in statistical analyses designed to understand teachers' performance and then separates out the variation associated with that content from final evaluations of teachers' performance, improvement, or response to feedback.

### Key Concept – Documentation/Data Collection

Record:

- Time of day
- Activity
- Adults and students present in the classroom
- Grouping
- Typicality of the observation period



*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.*