

Observation Checklists versus Observation Data

By Dr. John L. Tenny, developer of the Data-Based Observation Method and the eCOVE Classroom Observation Software, January, 2010

I have looked at a couple dozen books on how to conduct classroom/teacher observations and have downloaded another 30 or so observation forms used by various districts across the country. I am struck by the general nature of most of them, and by the lack of specifics in the descriptors. The rating scales used run from observed/not observed to met/not met to a 5 to 7 likert scale. Some of the forms used by districts are designed to record anecdotal notes without a rating scale and/or with the evaluation of the performance on another summary page.

An example: This 'standard' came from a school district set of guidelines for conducting observations: "Establishes and maintains an orderly and supportive environment for students", and is, in one form or another, common in standards.

I just cannot see how a checklist or scale is helpful, let alone accurate, as a record of what happened in the class. If the observer checks 'observed', does that mean the class was at one point orderly and supportive? Alternatively, does it mean that when things started to get disorderly, the teacher responded and brought the class back in focus? On the other hand, since the phrase 'supportive environment' is also included, does that 'observed' indicate that the teacher made positive, encouraging statements? To all the students? Could the observer be satisfied with student work being posted on the walls, and a 'student of the week' bulletin board being present -- that is certainly supportive?

In addition, if the class was orderly some of the time and not others, and the

observer checks 'not observed', would not the teacher respond with "Are you saying my class was never orderly Or that I was never supportive? Or both?"

Observed/ not observed does not work. It does not convey any helpful information and will lead to conflict between the observer and observee.

So what about a scale? Scales are typically designed to be either a 1 to 5 (poor to great) or a rubric of 'unsatisfactory, basic, emerging, competent, distinguished' type. Some of them will have descriptors for each of the levels, the worst being the range from 'did not observe/ observed some of the time/ observed most of the time/ observed all of the time'. These descriptors are worthless as they are nearly impossible to mark in a way that conveys what happened. For example, if the class were orderly for the first 3 minutes and then in chaos the rest of the time, the orderly standard would actually be met 'some of the time'. Actually, if the class were orderly up to 49% of the time, the same checkbox would apply; and if things were orderly 51% to 99% of the time, it would be 'most of the time'.

There are other descriptors or indicators for each of the levels that seem more specific. For example, Charlotte Danielson in her Framework for Teaching has a standard for Management of Instructional Groups with a proficient level of competence described as "Tasks for group work are organized, and groups are managed so most students are engaged at all times." In the many districts that use some variation of Danielson's work as their standards, the observer would be asked to judge if the teacher was at this level during the current observation.

Skip the fact that there are two behaviors in this indicator - organizing tasks and managing groups - which also confuses the issue, and just look at the act of

making that determination of worth based on the second behavior. When you watch a typical classroom, the complexity in deciding if 'most' (is that 51% or is it really a higher target than that?) 'are engaged' (physically or mentally? engaged in low level or high level work?) 'at all times' (what would be the determination if the full class went off task for 3 minutes?), has such wide variation across classrooms and observers that the validity is suspect.

In discussions with administrators, what I find very often happens is that the observer adds additional criteria to the specific situation. 'Most students' sometimes turns into almost everyone in the class (a higher standard than stated); 'engaged' equates to looking and acting busy without regard for the level or quality of engagement; and 'at all times' is ignored in lieu of an unspoken criteria of 'most of the time'. If the class is known to have kids with behavior problems or the number of students is high, the criteria is functionally lowered.

What is really happening is that the observer has internally defined a level that is satisfactory to him or her based on personal experiences, the makeup of the class, and the relationship with the teacher, and that definition is applied unevenly across classrooms. That inconsistency is confusing to everyone, and the results of classroom observations cannot be compiled across the building, let alone the district, as a basis for broad decisions. The system has a built in subjectivity and personal interpretation of the standards that makes it difficult for any observer to be consistent and fair.

Rating scales do not work, especially when extremely little effort is put into rater reliability and clear statement of the objectives and indicators.

Data based observations can make a significant difference. Some of the texts on classroom observations provide steps on how to turn a judgment on a rating scale into numbers and then process those numbers as if they were data -- but given all the issues with rating scales I believe this to be a false path.

Instead, I recommend using the Data-Based Observation Method, a 5-step process that includes the actual collection of observable behavior data. The steps are:

1. Identify the standards. Be sure that they are worded so that observable behaviors demonstrating those standards can be clearly identified.

Good standard: Students will be engaged in learning activities. Bad standard: Teachers will act in an ethical and professional manner at all times (what does 'ethical' look like?).

2. Create indicators. Be sure that they describe the observable behavior identified in the standard.

Good indicator: Students will listen attentively to the teacher, be productively engaged in individual work, or contribute to the work of a small group. Bad indicator: Students will follow teacher instructions as given (too vague and general).

3. Set criteria. As a profession, we have not engaged in setting criteria for ourselves in concrete terms, so this part is a new conversation.

What are the criteria for engaging students in learning? Should they be

engaged 25% of the time? No, that is clearly too low. How about 50% of the time -- still sounds low. What about 95% of the time? Too high for real classes? The answer here is not to set the criteria arbitrarily, but to turn to (or conduct) research to establish criteria in which we can have confidence.

If the standard is an important one, and the indicators are valid, there will be a correlation between the behavior observed (such as student engagement) and the final desired outcome (student learning). We need to find those connections and use them as guides for improving teaching. We actually have research that identifies a significant number of them, but we are not applying that research at the classroom level.

4. Design data collection tools. I developed the eCOVE Classroom Observation Software as an easy and efficient way to collect the objective data, but you can use pencil and paper, a stopwatch, the wall clock in the classroom, etc. to collect the data once you have carefully identified what data is important to collect.

Good tool: A counter tracking on/off task behavior and using the time sample data collection method to record the percent of time engaged and the percent of time not engaged for the entire class. By using the time sample data collection approach and repeated sweeps of the class to record the on/off task behavior of each student, a quite accurate data-based, objective picture of the class behavior is produced. This becomes a factual basis for making decisions. Useful tools include Class Learning Time, Level of Questions, Teacher Talk/Student Talk, and other tools reflecting

research on best teaching practices.

5. Analyze and interpret the data. Did it meet the criteria? Is there a need or desire for a change?

Given the context (number and diversity of the students, physical space, materials at hand, etc) what is most likely to bring a positive change? When and where will the new approach be initiated? When will the next set of data be collected, analyzed, and interpreted?

For the greatest success, it is critical to operate with the belief that every vested interest be involved in this process. Administrators, teachers, parents, aides, students, counselors, etc. all have an important contribution to make where the purpose of the observation is the improvement of teaching and learning.

I am coming to realize what a big shift this is in the education field. We have tried to cite 'professional judgment' when the inconsistency in the process and the unreliability of the results support neither the process nor the conclusions. Serious collaborative discussions are needed to move to a more concrete basis for judging what we say we value, and how to use the specifics to guide the improvement of teaching.