How Can Teachers Be Assured Trustworthy Results from Classroom Observations?[9]

Classroom observations can be powerful tools for professional growth. But for observations to be of value, they must reliably reflect what teachers do throughout the year, as opposed to the subjective impressions of a particular observer or some unusual aspect of a particular lesson. Teachers need to know they are being observed by the right people, with the right skills, and a sufficient number of times to produce trustworthy results. Given this, the challenge for school systems is to make the best use of resources to provide teachers with high-quality feedback to improve their practice.

"For the same total number of observations, incorporating additional observers increases reliability."

The MET project's report *Gathering Feedback for Teaching* showed the importance of averaging together multiple observations from multiple observers to boost reliability. Reliability represents the extent to which results reflect consistent aspects of a teacher's practice, as opposed to other factors such as observer judgment. We also stressed that observers must be well-trained and assessed for accuracy before they score teachers' lessons.

But there were many practical questions the MET project couldn't answer in its previous study. Among them:

- Can school administrators reliably assess the practice of teachers in their schools?
- Can additional observations by external observers not familiar with a teacher increase reliability?
- Must all observations involve viewing the entire lesson or can partial lessons be used to increase reliability? And,
- What is the incremental benefit of adding additional lessons and additional observers?

These questions came from our partners, teachers, and administrators in urban school districts. In response, with the help of a partner district, the Hillsborough County (Fla.) Public Schools, the MET project added a study of classroom observation

Hillsborough County's Classroom Observation Instrument

Like many school districts, Hillsborough County uses an evaluation instrument adapted from the Framework for Teaching, developed by Charlotte Danielson. The framework defines four levels of performance for specific competencies in four domains of practice. Two of those domains pertain to activities outside the classroom: Planning and Preparation, and Professional Responsibility. Observers rated teachers on the 10 competencies in the framework's two classroom-focused domains, as shown:

Domain 2: The Classroom Environment	Domain 3: Instruction
▪ Creating an Environment of Respect and Rapport	▪ Communicating with Students
▪ Establishing a Culture of Learning	▪ Using Discussion and Questioning Techniques
▪ Managing Classroom Procedures	▪ Engaging Students in Learning
▪ Managing Student Behavior	▪ Using Assessment in Instruction
▪ Organizing Physical Space	▪ Demonstrating Flexibility and Responsiveness

reliability. This study engaged district administrators and teacher experts to observe video-recorded lessons of 67 Hillsborough County teachers who agreed to participate.

Comparison of Ratings

Two types of observers took part in the study: Fifty-three were school-based administrators—either principals or assistant principals—and 76 were peer observers. The latter are district-based positions filled by teachers on leave from the classroom who are responsible for observing and providing feedback to teachers in multiple schools. In Hillsborough County's evaluation system, teachers are observed multiple times, formally and informally, by their administrators and by peer observers. Administrators and peers are trained and certified in the district's observation instrument, which is based on Charlotte Danielson's Framework for Teaching.

These observers each rated 24 lessons for us and produced more than 3,000 ratings that we could use to investigate our questions. MET project researchers were able to calculate reliability for many combinations of observers (administrator and peer), lessons (from 1 to 4), and observation duration (full lesson or 15 minutes). We were able to compare differences in the ratings given to teachers' lessons by their own and unknown administrators and between administrators and peers.

Effects on Reliability

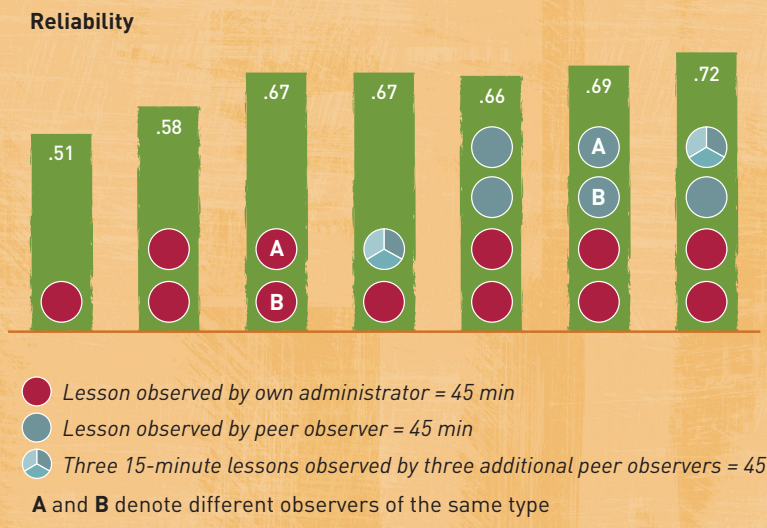
Figure 5 graphically represents many of the key findings from our analyses of those ratings. Shown are the estimated reliabilities for results from a given set of classroom observations. Reliability is expressed on a scale from 0 to 1. A higher number indicates that results are more attributable to the particular teacher as opposed to other factors such as the particular observer or lesson. When results for the same teachers vary from lesson to lesson or from observer to observer, then averaging teachers' ratings across multiple lessons or observers decreases the amount of "error" due to such factors, and it increases reliability.

Adding lessons and observers increases the reliability of classroom observations. In our estimates, if a teacher's results are based on two lessons, having the second lesson scored by a second observer can boost reliability significantly. This is shown in **Figure 5:** When the same administrator observes a second lesson, reliability increases from .51 to .58, but when the second lesson is observed by a different administrator from the same school, reliability increases more than twice as much, from .51 to .67. Whenever a given number of lessons was split between multiple observers, the reliability was greater than that achieved by a single observer. In other words, for the same total number of observations, incorporating additional observers increases reliability.

Of course, it would be a problem if school administrators and peer observers produced vastly different results for the same teachers. But we didn't find that to be the case. Although administrators gave higher scores to their own teachers, their rankings of their own teachers were similar to those produced by peer observers and administrators from other schools. This implies that administrators are seeing the same

Figure 5

There Are Many Roads to Reliability



Lesson observed by own administrator = 45 min
Lesson observed by peer observer = 45 min
Three 15-minute lessons observed by three additional peer observers = 45 min
A and **B** denote different observers of the same type

These bars show how the number of observations and observers affects reliability. Reliability represents the extent to which the variation in results reflects consistent aspects of a teacher's practice, as opposed to other factors such as differing observer judgments. Different colors represent different categories of observers. The "A" and "B" in column three show that ratings were averaged from two different own-school observers. Each circle represents approximately 45 minutes of observation time (a solid circle indicates one observation of that duration, while a circle split into three indicates three 15-minute observations by three observers). As shown, reliabilities of .66–.72 can be achieved in multiple ways, with different combinations of number of observers and observations. (For example, one observation by a teacher's administrator when combined with three short, 15-minute observations each by a different observer would produce a reliability of .67.)

things in the videos that others do, and they are not being swayed by personal biases.

If additional observations by additional observers are important, how can the time for those added observations be divided up to maximize the use of limited resources while assuring trustworthy results? This is an increasingly relevant question as more school systems make use of video in providing teachers with feedback on their practice. Assuming multiple videos for a teacher exist, an observer could use the same amount of time to watch one full lesson or two or three partial lessons. But to consider the latter, one would want to know whether partial-lesson observations increase reliability.

Our analysis from Hillsborough County showed observations based on the first 15 minutes of lessons were about 60 percent as reliable as full lesson observations, while requiring one-third as much observer time. Therefore, one way to increase reliability is to expose a given teacher's practice to multiple perspectives. Having three different observers each observe for 15 minutes may be a more economical way to improve reliability than having one additional observer sit in for 45 minutes. Our results also suggest that it is important to have at least one or two full-length observations, given that some aspects of teaching scored on the Framework for Teaching (Danielson's instrument) were frequently not observed during the first 15 minutes of class.

"Although administrators gave higher scores to their own teachers, their rankings of their own teachers were similar to those produced by external observers and administrators from other schools."

Together, these results provide a range of scenarios for achieving reliable classroom observations. There is a point where both additional observers and additional observations do little to reduce error. Reliability above 0.65 can be achieved with several configurations (see **Figure 5**).

Implications for Districts

Ultimately, districts must decide how to allocate time and resources to classroom observations. The answers to the questions of how many lessons, of what duration, and conducted by whom are informed by reliability considerations, as well as other relevant factors, such as novice teacher status, prior effectiveness ratings, and a district's overall professional development strategy.