Principals' use of rating scale categories in classroom observations for teacher evaluation

Stefanie A. Wind, Chia-Lin Tsai, Sara B. Grajeda, Christi Bergin
Published in *School Effectiveness and School Improvement*, Volume 29, Issue 3; pages 485-510
https://doi.org/10.1080/09243453.2018.1470989

Purpose of the study

The Network for Educator Effectiveness seeks to continue informing practices of educator evaluation. We do that through our work with schools, and we do that through our research. This article combines both efforts by analyzing how instructional leaders scored teaching episodes and the effectiveness of teaching practices during the recertification exam that each instructional leader took during the summer of 2015. The study focused on two questions:

1. Is the structure of a classroom observation rating scale comparable across teaching episodes and practices?
2. Do rating scale categories function as intended across teaching episodes and practices?

In other words, is the structure of the NEE scoring rubric comparable across various teaching episodes and teaching practices? And do the NEE scoring rubrics function as intended across various teaching episodes and teaching practices?

Background of the study

To complete this research, the article used data from all recertification qualifying exams from the Network for Educator Effectiveness Recertification Training during the Summer of 2015. In total, that consisted of 1,324 instructional leaders holding building- and district-level positions throughout Missouri.

The qualifying exam consisted of four teaching episodes. Teaching episodes are defined as video clips of approximately 10 minutes in length showing instruction of a teacher to a whole classroom of students. The four classrooms used for the exam were: a fourth grade mathematics classroom, a fifth grade language arts classroom, a ninth grade mathematics classroom, and a high school international baccalaureate classroom. All instructional leaders evaluated six teaching practices based on the scoring rubrics developed by the Network for Educator Effectiveness. The six indicators evaluated were:

- The teacher uses academic language to communicate key concepts of the discipline(s), and gets students to use academic language accurately.
- The teacher uses strategies to cognitively engage students.
- The teacher uses strategies to get students to problem solve and think critically.
- The teacher use motivation strategies effectively.
- The teacher has positive interactions or has a positive relationship with students.
- The teacher conductions on-going assessment of learner progress and understanding during the lesson.

The study looked for whether the NEE scoring rubrics were interpreted consistently across the four different classrooms and the six various teaching practices.

To complete the study, the researchers used the many-facet Rasch measurement theory to answer the research questions. Using a computer program, the researchers developed models to estimate teacher, evaluator, and rubric score on a linear scale. That scale provided insights into how principals evaluate a teacher's use of specific teaching practices during a classroom observation.

Findings of the study

Through the analysis, the researchers found 3 key results.

The first result they found was that that the rubrics were functioning as designed and that there was consistency. This result supports the notion that instructional leaders can accurately score classroom observations on specific teaching practices in various content areas, grade levels, and at various levels of effectiveness.

The second result was that the scores functioned as expected for three of the four classroom videos observed and that the scores functioned as expected for five of the six teaching practices observed. In both exceptions, there is evidence of inflation by instructional leaders. This finding of inflation in classroom observation is consistent with other research conducted by the Network for Educator Effectiveness, as well as research of other classroom observation protocols. Such inflation results in less measurement precision in the lower scores of the scoring rubrics. It also hinders the ability to accurately compare and interpret scores in a consistent manner.

The third result showed that not all scores are used across classroom observations or teaching practices. In initial findings of this research project, the researchers found that a score of 4 was not given as often as should be expected. Upon further analysis of classroom observations within the NEE database (24,000 teachers), they found a similar underuse of the score of 4 no matter the teacher, the observer, or the indicator being observed. This may happen because, within the NEE scoring rubrics, a score of 4 is at the midpoint of the rubric or because it does not have a concrete descriptor attached to it. However, scores of 2 and 6 are not underutilized in the same manner.

Reflection on the study

The analysis helps in validating the NEE classroom observation ratings and in identifying areas where the Network for Educator Effectiveness, and other evaluation systems, can continue to improve.

Consistent with other research, these findings suggest that even in well-developed classroom observation systems, scoring rubrics do not always function as intended and that the interpretation of the rubrics is not always consistent. In actual classroom observations, that could mean a percentage of classroom observation scores may not be validly interpreted.

However, this article also supports the notion that instructional leaders are able to consistently use the scoring rubrics to help define the teaching practices happening in a classroom. This happened across a wide range of content areas and grade levels and across six different specific teaching practices. Further training and time spent scoring classroom observations through guided and individual practice should continue, if not improve, that consistency.