Charlotte Danielson's Framework for Teaching

*Conducted by Steven Kimball and Anthony Milanowski.*

**Purpose of study**

To examine the validity of principals’ teaching evaluation ratings and determine whether differences in decision making contribute to the differential validity observed in these ratings. [Validity was defined by the relationship between evaluation ratings and value-added measures of student achievement.]

**Research questions**

* How much does the validity of the performance rating relationship vary across evaluators?
* Are differences in evaluator decision making in a standards-based teacher evaluation system related to differences in the strength of the student achievement- performance rating relationship?

**Population/sample**

The study was carried out in a large school district in the western United States, comprised of 88 schools in which about 3,300 teachers educate more than 60,000 students, that had implemented a teacher evaluation system based on Danielson’s (1996) Framework for Teaching several years prior and thus had performance evaluation results for many teachers over several consecutive years. After exclusions for missing evaluation scores and student achievement data, a total of 5,683 students and 328 teachers were included in the analysis for the 2001-02 school year, and 9,873 students and 569 teachers were included in 2002-03. In 2001-02, 39 principals evaluated 5 or more teachers and in 2002-03, 57 administrators evaluated 5 or more teachers. For the qualitative component of the study, 23 of the evaluators with more valid ratings (high correlations with teachers’ classroom average student achievement) and less valid ratings (very low and negative correlations with teachers’ classroom average student achievement) were selected for interviews. Additional analysis of interview transcripts and written feedback to teachers was conducted for eight evaluators with two consecutive years of high (average r=.55) or low (average r=-.28) validity ratings.

**Major results**

* Evaluators varied substantially in the strength and direction of the relationship between their teacher ratings and the achievement of those teachers’ students. Approximately 30% of administrators’ ratings of teachers had correlations with average student achievement below –.10 (low validity), while over 40% of administrators gave teachers ratings that correlated with their students’ achievement at .31 or higher (high validity).
* Differences in evaluators’ motivations, or will, to conduct teacher performance evaluations were not related to the validity of their ratings. All administrators reported positive attitudes about the evaluation system and emphasized its value for teacher development over accountability. Attitudes towards and perceived level of accuracy in teacher evaluations were also similar between high- and low-validity groups, despite comments that there was little oversight or consequences from the district. Reported compliance with the evaluation process was also high for both groups of evaluators, except for the supplemental evaluation form.
* Administrators in the higher validity group noted the value of district trainings in helping to build their evaluation skills, while only one of the lower validity evaluators talked about training from the district. Lower validity evaluators said that their experience teaching, in administration, or in business, was most helpful in conducting evaluations.
* Evaluators in the more valid group tended to talk about using rubrics in a more analytical way. However, administrators in each group mentioned factors that led them adjust ratings toward more lenient assessments of teachers.
* There were few differences in evaluator preparation between high- and low-validity groups, with most preparation focusing on preparing teachers for the process. Administrators in both groups tapped multiple, similar sources of evidence to conduct evaluations, most commonly classroom observations. Several teachers in both the high- and low-validity groups used a very structured approach to teacher evaluation.
* In terms of school environment, the socioeconomic status and prior achievement of students did not differentiate the high- from low-validity groups of evaluators. Neither did the teaching or administrative experience of the evaluators nor their perceived credibility to the teachers they evaluated. In all but one case, relationships between teachers and evaluators were positive.
* Evaluators overwhelmingly emphasized the formative purposes of teacher evaluation and tended to be lenient, focusing on praise in written evaluations rather than constructive feedback or suggestions for improvement. This leniency may reduce the validity of the evaluations.

**Conclusions/recommendations**

The authors conclude that evaluator will, skill and evaluation context did not have a clear or consistent influence on the validity of their teacher evaluations, that is, on the extent to which the evaluation scores were related to student achievement. Possible interpretations of this finding included complex, idiosyncratic interactions between evaluator characteristics, evaluators’ reliance on gut-level feelings about teachers, a lack of expectations or support to conduct evaluations accurately (e.g., low levels of accountability, no follow-up training, no consequences for most teachers), and a lack of sufficient detail in the data to reveal differences in evaluator will, skill and context.

Findings suggest that generating evaluation scores that are highly related to student achievement will take more than specific rubrics and basic training of evaluators. The authors advocate for the development of a “strong situation” that provides incentives for accurate evaluation, oversight that emphasizes ratings differentiate among teachers, and ongoing training and practice with feedback on accurate evaluations. They also caution against the use of teacher evaluation scores in making high-stakes decisions about tenure, promotion and compensation.

**FFT focus**

The teacher evaluation system in this district was based on the Framework, and the authors include detailed information about the instrument and its use in the article. The fact that some teachers’ scores on the Framework-based rubric and their students’ average achievement were highly correlated provides validation that the FFT describes observable practices that are associated with teaching that raises student achievement.