Title:

Improving Teacher Evaluation: Using Evidence from an RCT to Compare Ratings Used for Feedback to Ratings Used for Research

Authors:

Andrew Wayne, Jordan Rickles, Mike Garet, Seth Brown, Mengli Song

Background/Context:

A decade of federal, state, and district efforts to reform teacher evaluation systems has produced numerous opportunities for researchers to study evaluation system practices. Much of the discussion has focused on the performance ratings that evaluation systems generate and the classroom practice ratings that undergird them. For example, many reformers point out that even though research shows wide variation across teachers in their contributions to student achievement, almost all teachers receive performance ratings labeled satisfactory or better (e.g., Kraft and Gilmour, 2017; Weisberg et al., 2008).

Purpose:

In this context, we examine a unique data set, derived from a recently completed cause-and-effect study (Garet et al 2017). Specifically, we compare parallel sets of ratings of teacher classroom practice, collected during the course of the study by different observers with different purposes and modes of observation. The purpose of this paper is to explore the conditions under which classroom practice measures produce ratings with the features that reformers and researchers have said they ought to have, including that these ratings should (1) correlate with student learning, (2) differ across teachers, and (3) indicate a teacher's strengths and weaknesses.

Setting:

The data come from a randomized controlled trial on the impact of providing supplemental performance feedback to teachers and principals (see Garet et al 2017). Of the 127 elementary and middle schools that participated, half of the schools in each of the eight districts were assigned to receive the supplemental feedback (treatment), which lasted for two years (2012-13 and 2013-14), and half to continue with business as usual (control). In each participating school, the study focused on the principal and teachers responsible for teaching reading/ELA and mathematics in grades 4–8. In the treatment group, the principal received feedback on his or her leadership, and teachers received feedback on their classroom practice and value-added.

Research Design & Population/Participants/Subjects:

The study generated two parallel sets of classroom practice data that serve as the basis for this article. One set—*the feedback ratings*—was the feedback to teachers on their classroom practice, generated by observers in the treatments schools only (n = 63 schools). The other

set—*the video ratings*—was collected in all schools (n = 127 schools) to assess the impact of the intervention on classroom practice independently, using video that was scored by the research team. The latter set was for use only by the researchers and not shared with the teacher or anyone outside the research team. One might expect the features of the feedback and video ratings to differ because of the different purposes of the raters (to give as feedback, or use among the researchers only) and other differences in methods.

An additional feature of the data set is that it allows us to compare these two sets of ratings (i.e., compare a teacher's feedback and video ratings) separately for two groups of teachers: one group was rated using the Classroom Assessment and Scoring System (CLASS) and the other using the Framework for Teaching (FFT).

The sample includes the 422 treatment teachers from the spring of the second year of the study who had at least one feedback rating and one video rating. Of them, 228 were rated using the CLASS and 194 using the FFT.

Data Collection and Analysis:

The paper discusses how the ratings were generated in detail (e.g., purpose of rating, mode of observation; see Exhibit 1).

In reporting the findings, we discuss specific types of scores. For a given observation, the CLASS and FFT instruments each require the rater to provide a set of **dimension scores**, consisting of a score of 1-7 for CLASS for each of its 10 dimensions, or 1-4 for FFT for each of 10 dimensions (see Exhibit 2 – map of domains and dimensions of CLASS and FFT). The **overall score** for an observation is the mean of the dimension scores. Finally, we compute the mean of the overall scores from across the spring of the second year, which we call the **spring-average overall score.** Likewise, a teacher's **spring-average dimension score** is the average of dimension scores from across observations.

We use these scores to compare features of the feedback and video ratings, using CLASS and FFT, focusing on whether the ratings have three features reformers and researchers have said they ought to have.

Findings/Results:

Feature 1: Classroom Observation Scores Should Correlate with Student Learning

We find that the spring average overall scores for feedback and video each generally correlated with value-added (VA) positively but weakly. Although the correlation between spring-average FFT overall score and VA in ELA/reading was negative, the other seven correlations estimated were positive, as shown in Exhibit 3.

Feature 2: Ratings Should Vary Across Teachers

For both CLASS and FFT, the distributions of the feedback ratings and video ratings both vary across teachers, with a similar standard deviation, but the feedback ratings are higher than the video ratings (Exhibits 4 and 5).

Feature 3. The ratings should indicate a teacher's strengths and weaknesses

With 10 dimensions on CLASS and FFT, we expect that ratings will indicate each teacher's strengths and weaknesses. That is, the spring-average dimension scores a teacher receives from an observation should differ. We found that for both CLASS and FFT, the difference between a teacher's spring-average dimension scores was greater when using the video ratings compared to the feedback ratings (Exhibits 6 and 7).

Conclusion. The full paper discusses the implications of these findings for efforts to generate useful ratings based on classroom observations. In an earlier study, principals gave higher ratings when the ratings counted in the evaluation system, compared to ratings for formative purpose. In this study, observers gave higher ratings when the ratings were for formative purposes, compared to ratings by other observers that were shared only with researchers. In other words, requiring observers to provide the ratings to teachers seems to have led to higher ratings. This paper also provides the first evidence on comparing the dimension ratings a teacher receives based on different observation methods.

References

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science, 333*(6045), 1034–1037.

Chaplin, D., Gill, B., Thompkins, A., and Miller, H. (2014). *Professional Practice, Student Surveys, and Value-Added: Multiple Measures of Teacher Effectiveness in the Pittsburgh Public Schools (REL 2014-024).* U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Washington, DC: Regional Educational Laboratory Mid-Atlantic.

Garet, M.S., Wayne, A.J., Brown, S., Rickles, J., Song, M., and Manzeske, D. (2017). *The Impact of Providing Performance Feedback to Teachers and Principals*, *Executive Summary* (NCEE 2018-4000). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Grissom, J.A., & Loeb, S. (2017). Assessing Principals' Assessments: Subjective Evaluations of Teacher Effectiveness in Low- and High-Stakes Environments. *Education Finance and Policy*.

Kane, T.J., Taylor, E.S., Tyler, J.H., and Wooten, A.L. (2011). Identifying Effective Classroom Practices Using Student Achievement Data. *Journal of Human Resources, 46*(3): 587– 613.

Kane, T.J., and Staiger, D.O. (2012). *Gathering Feedback for Teaching: Combining High Quality Observations With Student Surveys and Achievement Gains.* Seattle, WA: Bill & Melinda Gates Foundation.

Kraft M.A., Gilmour A.F. (2017). Revisiting the Widget Effect: Teacher Evaluation Reforms and the Distribution of Teacher Effectiveness. *Educational Researcher, 46*(5): 234-249.

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunk, J., Palcisco, A., & Morgan, K. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness* (2nd ed.). The New Teacher Project.

Exhibit 1: Key aspects of the collection of the feedback ratings and the video ratings.

Similarities and Differences	Feedback ratings	Video ratings
Purpose of rating	Formative feedback to the teacher	An outcome measure for the impact study, not shared outside the research team
Mode of observation	In-person	Video collected by local videographer and coded at research firm
Type of rater	A trained observer from another school or hired locally	Research staff
Number of observations in spring of second intervention year	1 or 2 times per teacher	1 or 2 times per teacher

Exhibit 2: Domains and dimensions of classroom practice measured by the Framework for Teaching (FFT) and the Classroom Assessment and Scoring System-Secondary (CLASS)

FFT[a]	CLASS
Domain 2: Classroom Environment • Creating an environment of respect and rapport • Establishing a culture for learning • Managing classroom procedures • Managing student behavior • Organizing physical space Domain 3: Instruction • Communicating with students • Using questioning and discussion techniques • Engaging students in learning • Using assessment in instruction • Demonstrating flexibility and responsiveness	Domain 1: Emotional Support • Positive climate • Negative climate • Teacher sensitivity • Regard for adolescent perspectives Domain 2: Classroom Organization • Behavior management • Productivity • Instructional learning formats Domain 3: Instructional Support • Content understanding • Analysis and problem solving • Quality of feedback Domain 4: Student Engagement • Student engagement

[a] The 10 FFT dimensions listed in this table are dimensions of classroom practice related to classroom environment and instruction that can be assessed through classroom observation. The full FFT instrument includes 12 additional dimensions related to lesson planning and preparation and professional responsibilities, which were not be measured in Garet et al (2017) because they are not readily amenable to classroom observation.

Additional notes on this exhibit: The CLASS and FFT have many similar features. For example, both focus on multiple dimensions of instruction, and the rating levels on each dimension are defined using specific, observable behaviors of teachers and students. There is evidence of validity and an association with student achievement for both instruments (Allen et al. 2011; Kane and Staiger 2012).

Exhibit 3: Correlation of the spring-average overall scores for classroom practice with value-added in reading/ELA and mathematics, for teachers rated using CLASS and FFT

	Correlation with spring-average overall score from video ratings	Correlation with spring-average overall score from feedback ratings	
Correlation with VA in reading/ELA -CLASS -FFT	 .06 -.03	 .12 .19	
Correlation with VA in math -CLASS -FFT	 .14 .15	 .13 .24	

Additional notes on exhibit: Although the positive correlations between classroom observation overall scores and value-added scores were modest in magnitude, these correlations are consistent with the magnitudes found by other studies (Chaplin et al. 2014; Kane and Staiger 2012; Kane et al. 2011) and likely underestimate the strength of the true association because of measurement error in both the observation scores and the value-added scores.

Exhibit 4: Distributions of spring-average overall scores based on the feedback (intervention) ratings and video ratings, for teachers rated using CLASS

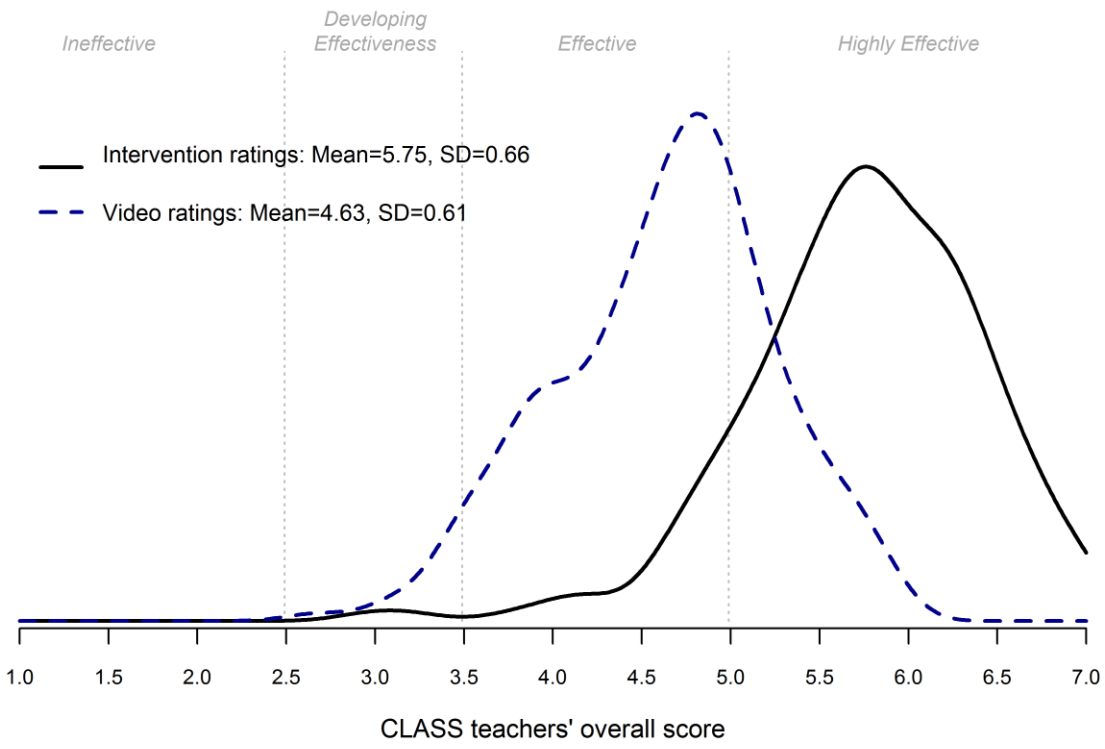

Exhibit 5: Distributions of spring-average overall scores based on the feedback (intervention) ratings and video ratings, for teachers rated using FFT

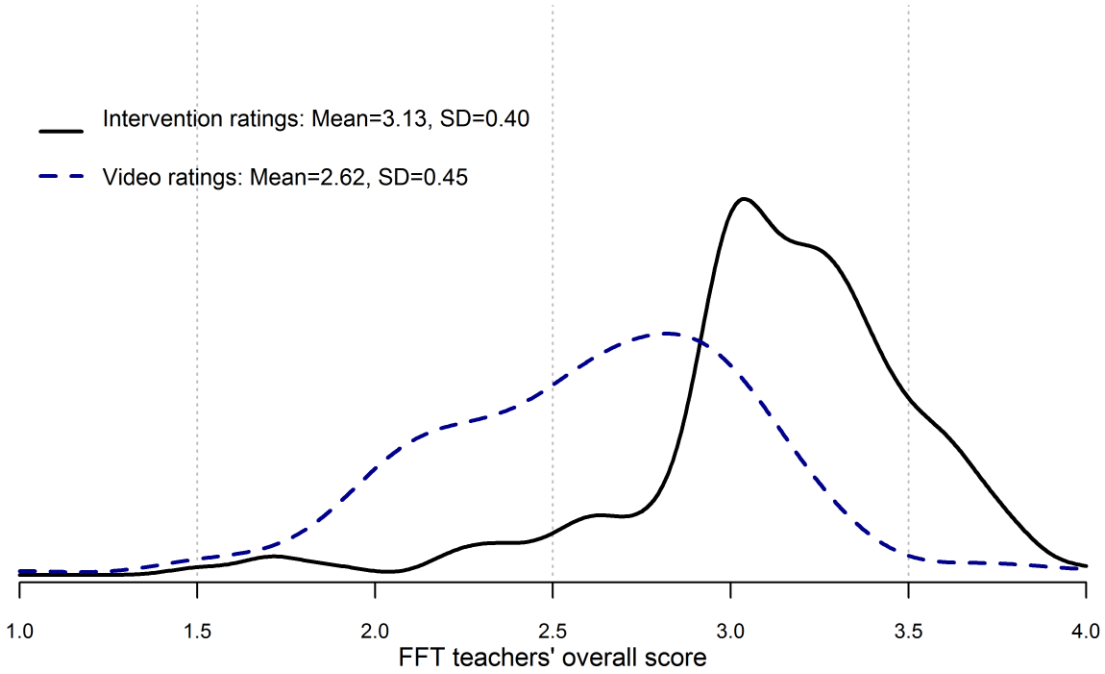

Exhibit 6: Distribution of teachers by number of scale points between their lowest and highest spring-average dimension scores, for teachers rated using CLASS

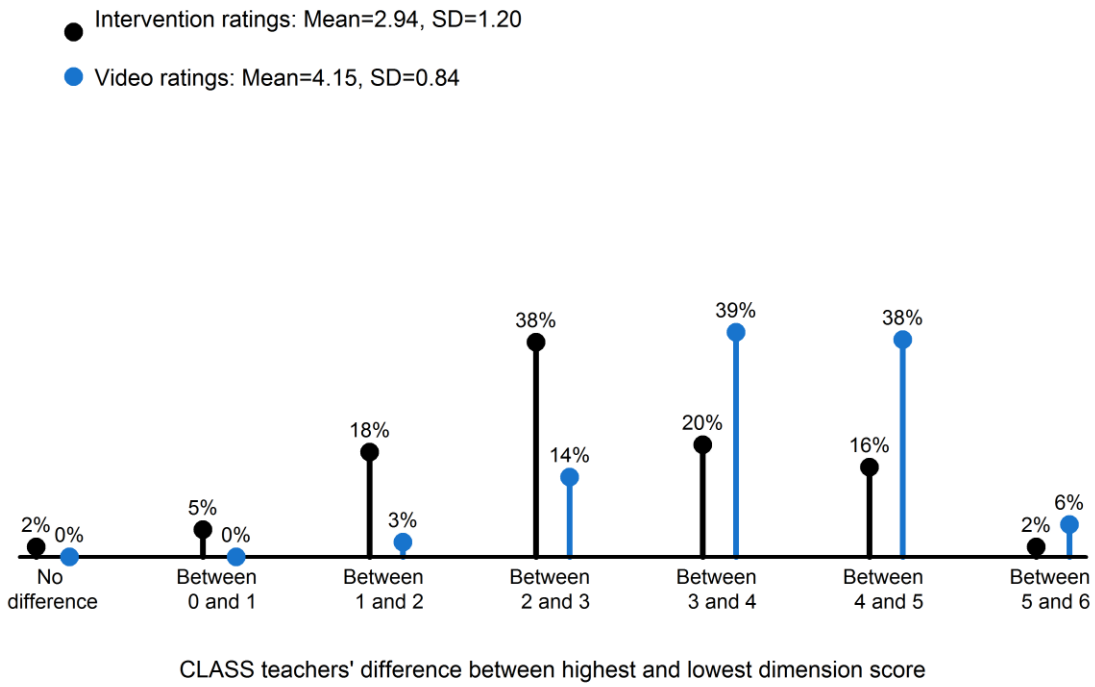

Exhibit 7: Distribution of teachers by number of scale points between their lowest and highest spring-average dimension scores, for teachers rated using FFT