Want Principals to Rate Teachers Honestly? Take Away the Stakes.

By Liana Loewus on July 7, 2017 4:26 PM

Principals are more likely to rate teachers as ineffective in a confidential setting than on a formal evaluation with stakes attached, a new study finds.

That's in part because principals want to maintain good relationships with their teachers, which can be tough to do when they have to confront them with bad scores, the researchers say.

Researchers have long flagged the lack of variation in teachers' evaluation ratings. In **2009, TNTP published its eye-opening report The Widget Effect**, which found that less than 1 percent of teachers were being rated as unsatisfactory. Many states began taking steps to **implement more rigorous evaluation systems**. The federal Race to the Top program, starting at the end of that year, offered states incentives to incorporate student test scores into those evaluations.

But teachers have **continued to get high—and potentially inflated—ratings**, the **research shows**.

"We've invested a lot in making these systems rigorous and yet they still seem to identify the vast majority of teachers as effective, especially when you look at the observation ratings from principals," said Jason Grissom, an associate professor of public policy and education at Vanderbilt University, who co-authored the study.

Low Stakes vs. High Stakes

Published in the journal Education Finance and Policy, the recent study analyzed how 100 principals from Miami-Dade County public schools rated the same teachers in two different settings: one low-stakes and one high-stakes. The researchers also compared those ratings to the teachers' value-added scores, which use student test data to measure how a teacher is doing.

For the low-stakes evaluation, the researchers sat down with principals and asked them about three or four of their teachers. What are the teachers' strengths and weaknesses? How do they contribute to the school environment? They were also asked to rate the teachers on a scale of 1 (very ineffective) to 6 (very effective) on a variety of in-class and out-of-class measures.

The results of that assessment, the principals understood, would remain confidential and would not be used for any purposes outside the research.

The researchers then looked at how the principals rated those same teachers several weeks later on the district-required evaluation—which does have stakes attached. Districts use such

evaluations to make decisions about compensation, performance-improvement plans, and dismissals.

"As it turns out, they give pretty different assessments in the two environments," said Grissom.

Overall, the ratings still trended positively in both settings. But principals were much more likely to use the "ineffective" categories in the low-stakes environment than on the formal district evaluation. Principals almost never used those categories on the high-stakes assessment. (Compare the chart to the right, on the high-stakes results, to the low-stakes distributions illustrated in the chart at bottom.)

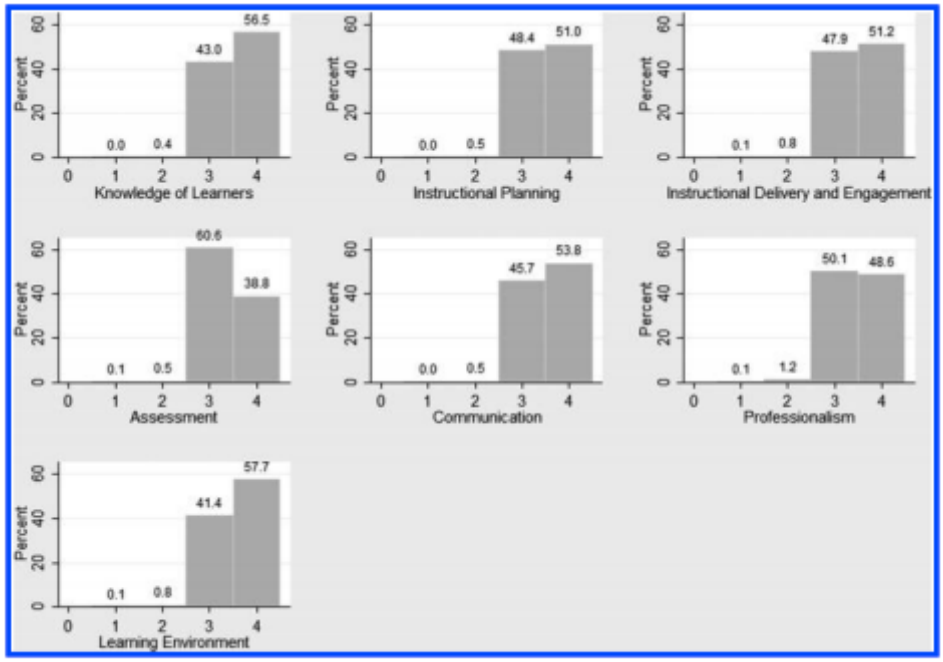

Figure 1. Distribution of Scores on High-Stakes Evaluation Instrument.

In fact, the teachers who received scores of "very ineffective" on the low-stakes assessment, on average were deemed "effective" on the high-stakes evaluation.

Uncomfortable Conversations

An important caveat here is that the high-stakes district evaluation used only a four-point scale (from 1, very ineffective, to 4, very effective), while the researchers' scale was six points. And the two evaluations asked slightly different questions, but about the same general themes, said Grissom.

It's possible that the differences in scales led to some of the differences in ratings. "A mistake we probably make in high-stakes evaluation systems is that we don't give raters enough categories," said Grissom. "Four categories is very constraining." (It's worth noting when the feds required states to make teacher-evaluation changes as a condition of receiving waivers from the now-defunct No Child Left Behind Act, they required only three categories.)

But that likely wasn't the main reason the scores diverged, he said.

With the district evaluations, "teachers know what the rating is," Grissom said. "In many systems, that involves a post conference. If I gave you low ratings, that would be very uncomfortable for

me to talk to you about. I want to maintain good relationships with my teachers; I want them to like me."

And of course there are the potential job consequences for teachers.

"It would be a rational response for a principal to think, if I give this person a low score, they might get angry and leave my school, or they might be dismissed, and then I have to replace this person and I might be facing a hiring pool that doesn't look appreciatively better than the teacher who would leave," he said. "Principals have a lot of reasons to evaluate the way they do."

Differentiation to a Degree

Interestingly, a closer look at the scores showed that the principals actually *were* differentiating between teachers in the high-stakes system. Even though nearly all teachers got 3s and 4s, both of which labeled them "effective," the principals seemed to be using those categories systematically.

"It appears principals are giving 4s to teachers they really think are the strongest, and giving 3s to teachers they think are less effective," said Grissom.

Teachers who got 3s on the district evaluation had lower scores on the confidential evaluation as well.

And lower scores on both of the evaluations correlated with lower value-added scores.

Principals "are capable of differentiating, but they also face really strong incentives to not fully differentiate when they know there are potential job consequences for their teachers or consequences for their own relationships with their teachers," Grissom said. "If policymakers want to see more differentiation, they have to take those incentives principals have very seriously."

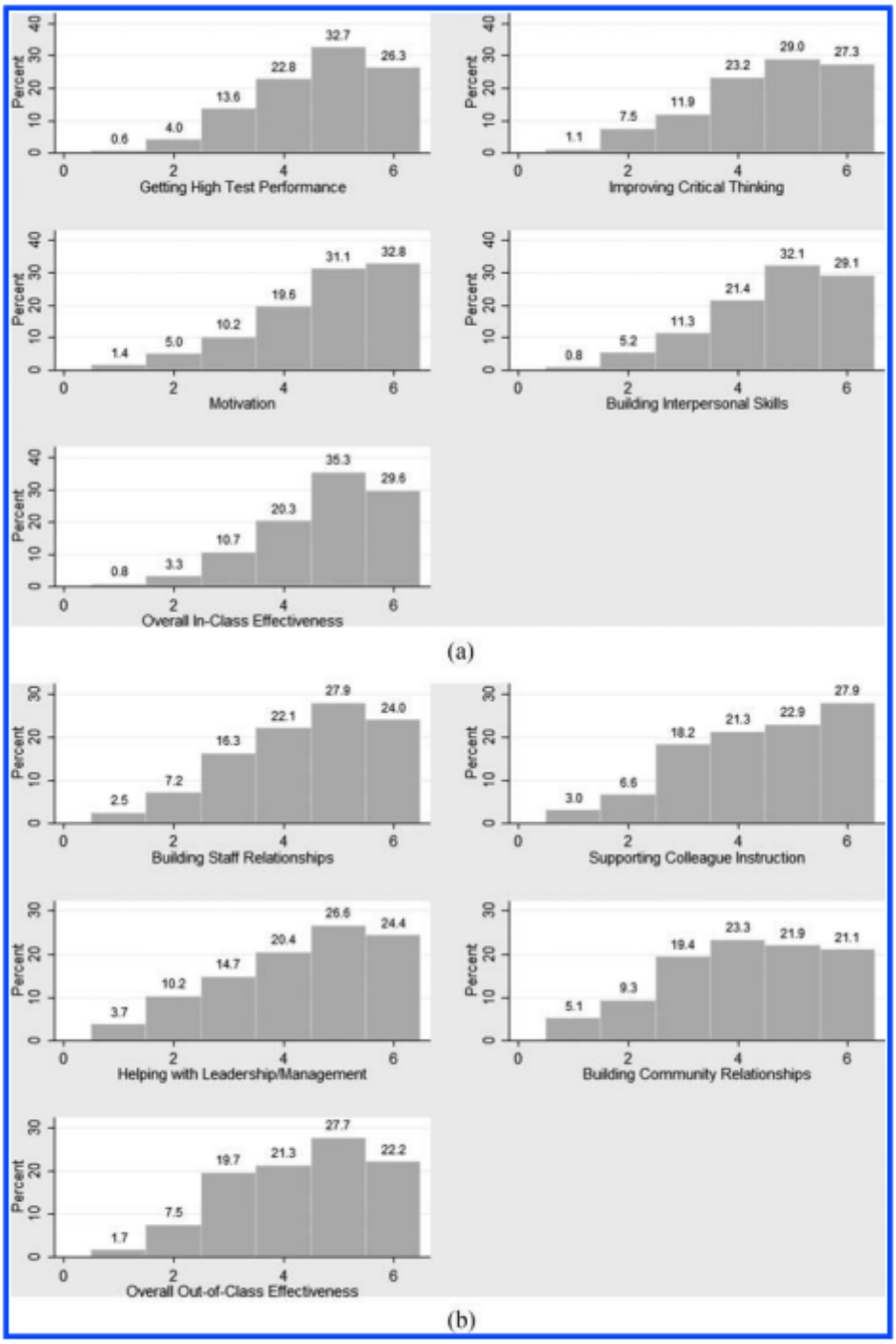

Figure 2. Distribution of Scores from Low-Stakes Interview Ratings. a. Distribution of Teacher Ratings on In-Class Items. b. Distribution of Teacher Ratings on Out-of-Class Items.

Charts: Education Finance and Policy