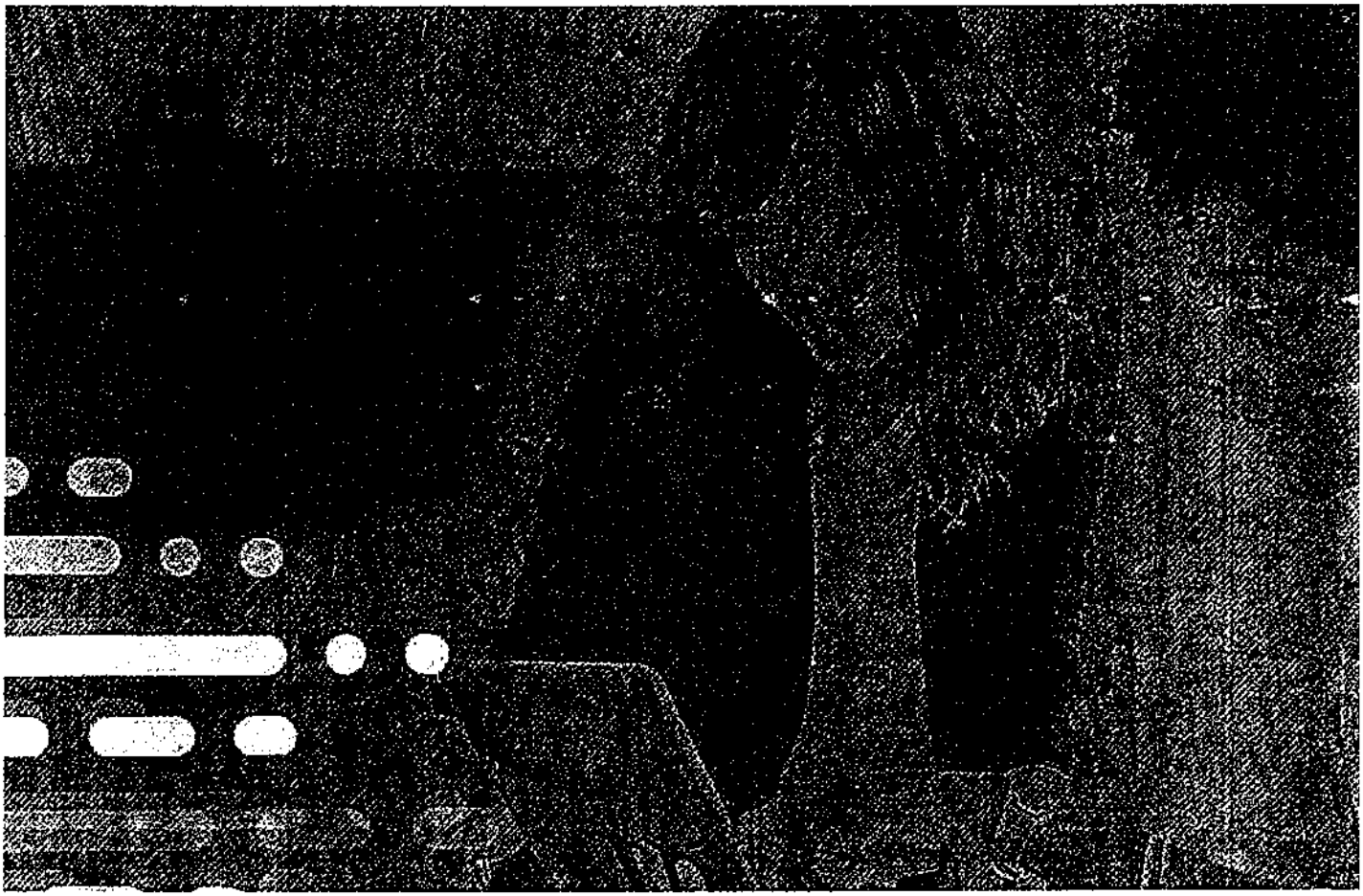
5 Tips to Understanding and Avoiding Bias in Teacher Performance Evaluations

PROFESSIONAL GROWTH By Xuanxuan Xu on 3/18/2018



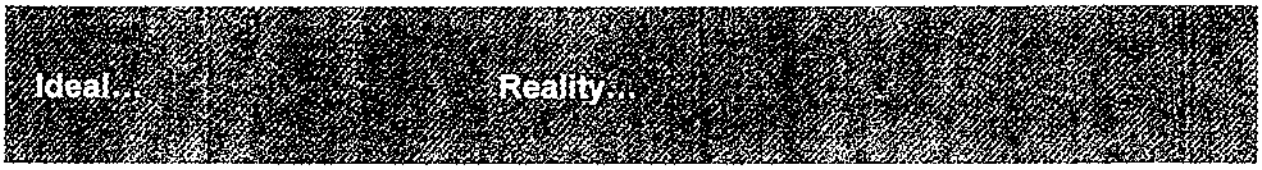
We are well into the last quarter of the school year, and that means we are entering the final phase of conducting classroom observations for teacher performance evaluation. If observations involve only one person, how do we enhance intra-rater reliability? And especially, if multiple observers have been

involved in observing teachers' classroom performance, what can we do to ensure inter-rater reliability? If a vice principal conducted an observation earlier in the year and the principal is scheduled to conduct an additional or final observation, how can we ensure the results of the more recent observation won't sway the decision disproportionately and that observers are using the same yardstick to measure effectiveness?

However unintentional, as evaluators, our biases can interfere with the accuracy and reliability of evaluations. Teachers won't trust and use evaluation results for improvement unless they are convinced the observations are accurate, truthful and justifiable.

"People only see what they are prepared to see." —Ralph Waldo Emerson

Error in evaluation, or in any measurement, is inevitable. Human performance – including teaching – is especially elusive to measure from the very beginning, and the validity of the evaluators is compromised by a number of factors (Bejar, 2012).



The scores evaluators assign to teachers only reflect:

- The true quality of performance

The scores evaluators assign to teachers also depend on:

- The quality of evaluators' understanding of the performance rubric

- The quality of the evaluators' *interpretation* of teachers' performance
- Fatigue and other factors that can influence evaluators
- Environmental conditions
- The nature of performance previously scored

Bias is normal and universal. We all perceive the world differently, and interpret what we observe differently. Our experiences shape our views and vice versa. Although we cannot be free of bias, if we can acknowledge and understand our biases, we will be better able to overcome their effects.

Five common bias issues in teacher observation and evaluation (and proposed solutions): http://bit.ly/2IpXz5t

CLICK TO TWEET THIS

Here are five common bias issues in teacher observation and evaluation, and proposed solutions to overcome them:

Issue 1: Rater Personal Bias

This bias occurs when evaluators apply idiosyncratic criteria that are irrelevant to actual teacher performance. Often without realizing it, evaluators give higher ratings to teachers who resemble them or have characteristics in common with them — for instance, certain beliefs or ways of getting things done which are not essential to an educator's effectiveness.

Likewise, the evaluator might give a lower rating because the teacher has different preferences for instruction, even though the instructional delivery is effective. If the teacher is rated too high or low based on a rater's personal bias, she will not know how to improve her teaching because she will not understand from where the rating really came.

Examples:

- *"This teacher reminds me of myself when I started teaching, so I'll give him a higher rating."*
- *"The teacher moves around a lot. I prefer to be more stationary when I teach. I'll give her a lower rating."*

Solution: Training evaluators on objective ways to collect evidence from multiple sources on uniform, research-based performance standards will help overcome this bias. When evaluators let their own judgments get in their way of accurately evaluating teachers, training can help them be more objective.

Issue 2: Halo and Pitchfork Effect

The halo and pitchfork effect can arise when early impressions of the educator being evaluated influence subsequent ratings. In the halo effect, this impression tends to be one that is too favorable. For example, let's assume a principal has a positive impression of a teacher who is professionally dressed. Even if the actual observation of the teacher suggests deficiencies in the teacher's performance, the evaluator might use more leniency than with other teachers who may not be dressed as professionally.

On the other hand, if the evaluator has a *negative* impression of a classroom where students scramble around the room and chat noisily minutes before the

lesson starts, the evaluator might then have less tolerance for that teacher, even when students are on task and engaged when the lesson begins. This would be an example of the pitchfork effect.

It may seem that the halo effect could help teachers, but if their ratings are inflated, they will not know how to improve and develop their instructional skills. If they are unjustly deflated they may get disheartened because they don't feel they are getting a fair assessment on their true abilities.

Examples:

- *"You were very professionally dressed and well-spoken, so I'll give you the benefit of the doubt if I see deficiencies in your classroom."* (Halo)
- *"The kids were rowdy and noisy as the lesson started, so now I look for flaws when I observe you."* (Pitchfork)

Solution: To help counter this issue, evaluators should be trained on objective ways to collect evidence on uniform, research-based criteria. Multiple evaluators also should be used, so that various perspectives are included. These solutions will help prevent an evaluator from rating a teacher inaccurately based on his or her own impressions.

Issue 3: Error of Central Tendency

Central tendency is a bias in which evaluators tend to rate all teachers near the middle of the scale and avoid extreme scores, even when such scores are warranted. This is a very common issue and can happen for various reasons: a desire to avoid hurting anyone's feelings, for example, or the worry that teachers will be upset if they realize their ratings are different.

If everyone receives the same rating, improvement is difficult. It discourages those teachers who are performing at a highly effective level, while giving false confidence to those who need significant improvement because their current performance is not meeting student needs. Essentially, the rating can perpetuate ineffective teaching practices.

Examples:

- *"We're all the same…and we're all acceptable!"*
- *"I don't want to upset anybody, so I am not going to differentiate and am going to rate everyone in the middle."*

Solution: One solution is to train evaluators to distinguish between the various ratings on the scale. Evaluators also should be trained on using precise feedback based on data-generated evidence. This is done formatively so the teacher can continually improve. These solutions help teachers receive accurate, helpful ratings rather than always being rated in the middle.

Issue 4: Error of Leniency

When evaluators tend to assign high ratings to a large sector of teachers when the ratings are not earned, this is known as leniency error. For instance, they might rate all or most of their teachers as highly effective, even when teaching performance or student growth and achievement measures do not justify these ratings.

While the reasons for this particular error are often well-meant, it causes similar problems as the error of central tendency. Leniency can frustrate high-performing teachers and keep lower-performing teachers from receiving the support they need to improve.

Examples:

- *"Everyone is superior…or better!"*
- *"We are living in the fictional town of Lake Wobegon, where everyone is above average!"*

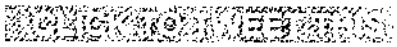
Solution: Train evaluators on distinguishing between the various rating levels so they can score teacher performance based on pre-defined criteria and the actual evidence collected. Evaluators likely rate teachers too highly when they do not clearly understand the differences between the ratings. Extra training will help them see the difference between effective and highly effective.

Issue 5: Rater Drift

With this, evaluators begin with a level of agreement on observations and ratings, but then gradually drift apart as they begin to apply their own spin to various criteria. Rater drift can happen at a collective level. For example, all evaluators might initially agree on what "student engagement" means, but over time come to define it differently. One evaluator might start to base it on how many students are looking at the teacher, while another looks at how many questions students ask and answer, and yet another focuses on student work from the lesson.

Rater drift also can happen to evaluators individually. A 2015 study by Casabianca and colleagues examined the ratings given to teachers based on observations. In the beginning, raters gave high scores. As time went on, however, they issued lower scores, even for the same teaching quality, ultimately dropping from about the 84th percentile to the 43rd — despite the fact that a teacher's quality had not changed.

"Rater Drift" can happen at the collective or individual level. Some examples here: http://bit.ly/2lpXz5t



Examples:

- *"Although my co-evaluators and I were trained and calibrated at the beginning of the school year, I am going to add my own personal twists down the road!"*
- *"I just read an interesting article about classroom management, and that changed my view of what productive classroom environment should look like. I will redefine the evaluation criteria!"*

Solution: This bias can be addressed by providing refresher training for evaluators and by using tandem reviews to ensure that evaluators are seeing things in the same way, making them less likely to drift away from each other in their ratings.

Bias and errors crop up when evaluators accidentally or habitually overlook, misinterpret, or distort what is perceived. Bias and errors confound the quality of evaluation, and that is why research-based calibration training is essential – training that prepares evaluators to know:

1) what effectiveness truly looks like and what to look for,

2) how to document teacher performance with objective evidence, and

3) how to synthesize evidence and apply the rubrics to provide ratings.

A solid training plan involves more than a one-shot calibration dose at the beginning of the school year. It also needs ongoing refresher training sessions on a recurring basis to make sure that evaluators consistently and persistently follow the prescribed criteria.

How can you ensure your evaluators and observers are trained and calibrated to provide reliable and defensible evaluations?

Learn about the Stronge Master-Coded Simulations and the Stronge Effectiveness Performance Evaluation System, powered by Frontline Professional Growth.

References:

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice, 31*(3), 2-9.

Casabianca, J. M., Lockwood, J. R., & McCaffret, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement, 75*(2), 311-337

This post was collaboratively authored by Xianxuan Xu, Ph. D. and Dr. James Stronge, Ph.D., President of Stronge and Associates Educational Consulting, LLC.

PAGE INTENTIONALLY BLANK