EdPolicyForward

The Unintended Effects of Policy-Assigned Teacher Observations on Observation Scores

By Dr. Seth Hunter and April Ege

April 30, 2019

Until the mid-2000s, the number of formal classroom observations assigned to teachers by policy was largely based on teacher years of experience or tenure (Steinberg & Donaldson, 2016). Now, the policy-based assignment of observations in several states is determined by teacher effectiveness, with less effective teachers assigned more observations. At face value, this change in observation assignment is laudable because students taught by more effective teachers experience better short- and long-run cognitive and non-cognitive outcomes (Chetty, Friedman, & Rockoff, 2014; Jackson, 2018). And, recent studies suggest that, under certain conditions, the receipt of more observations improves teacher effectiveness (Hunter, 2019a; Papay & Richard, 2017), although there is also strong evidence to the contrary (de Barros, 2019; Hunter, 2019b).

Yet, the number of observations assigned to teachers by state policy may (un)consciously lead observers to generate biased observation scores. Observers may believe teachers assigned more observations are worse teachers, and (un)wittingly issue lower observation scores, independent of observed teacher performance. Such bias will also bias composite measures of teacher effectiveness that heavily rely on teacher observation scores, as most composites do (Steinberg & Kraft, 2017). To the extent teacher professional development and personnel policies (e.g., retention, bonus pay) depend on composite measures of teacher effectiveness, resources may be substantially misallocated. Therefore, sources of bias in teacher observation scores are of interest to both policymakers and practitioners.

Strong Evidence of Negative Bias at Thresholds Where State Policy Assigns Observations

Using three years of teacher panel data from more than 80 percent of Tennessee districts, a working paper by EdPolicyForward's Seth Hunter finds strong suggestive evidence that relatively (in)effective teachers receive substantially lower observation scores when state policy assigns them more observations

EdPolicyForward
The Center for Education Policy @George Mason University

(Hunter, 2019c). Critically, the evidence suggests it is the assignment of observations, not the receipt, that introduces this bias. For example, a relatively (in)effective teacher assigned to receive two instead of one policy-assigned observations over the course of a year is predicted to receive substantially lower observation scores on her first observation of the year. Yet, there is no way the receipt of a teacher's second observation can genuinely affect her first observation score, suggesting the negative relationship captures some form of observer bias. At the beginning of each school year, observers learn how many policy-assigned observations teachers should receive. This observer knowledge may (un)intentionally introduce observer bias into first observation scores. Among relatively ineffective teachers, the assignment of an additional observation by state policy is predicted to lower first observation scores, but all predictions are statistically insignificant and are not discussed any further. However, when state policy assigns relatively effective teachers an additional observation, their first observation scores are predicted to decline by a statistically significant 0.2 standard deviations. This decline is equivalent to approximately half the difference between the average observation scores of first- and second-year teachers, representing a substantial decline.

Dr. Hunter estimates these relationships using a "regression discontinuity design," one of the strongest research designs following a randomized control trial (Shadish, Cook, & Campbell, 2002). Tennessee policy assigns observations based on whether a teacher is above or below certain thresholds on its composite measure of teacher effectiveness. The research design compares teachers just to either side of each threshold within small bandwidths, treating whether a teacher is located just to either side of a threshold as a process of "local randomization." Because teachers in these small bandwidths are "locally randomized," and because falling to one or the other side results in a different number of policy-assigned observations, the research design mimics a process by which teachers are randomly assigned some number of observations by state policy. However, a potential drawback of this research design is limited generalizability: findings may not generalize beyond the bandwidths surrounding thresholds.

Ancillary analyses explore confounding influences, with the most plausible being that the assignment of more observations by state policy to relatively effective teachers is confounded with teacher assignment to a lower category of effectiveness. That is, when relatively effective teachers cross from above to below the threshold they are assigned to a lower category of effectiveness and more observations by state policy, either of which may introduce bias. Analyses strongly suggest that the assignment of a relatively effective teacher to a lower category of effectiveness is

not the source of bias, although this suggestion is not definitive.

Implications

The evidence leads to three implications: two radical and one realistic. First, policy could assign all teachers the same number of observations. But this would almost certainly be an unproductive solution. If all teachers were assigned the same number as the least effective teachers this would increase the administrative burdens of teacher evaluation, which administrators already report as being quite burdensome (Kraft & Gilmour, 2016; Rigby, 2015). Alternatively, if all teachers were assigned the same number as the most effective teachers, this may result in foregone opportunities to improve the effectiveness of the least effective teachers via observational processes. Second, policymakers could remove observation scores as a high-stakes measure of teacher performance, reducing the importance of observer bias. However, the recent trend has been for states to give increasingly more weight to observation scores (e.g., Tennessee Board of Education, 2013; Tennessee Department of Education, 2016).

The most realistic course of action in response to this work may be for policymakers and practitioners to mitigate observer bias. Some work suggests intensive, and ongoing professional development for observers can reduce large degrees of observer bias (Milanowski, 2017). However, the effectiveness of such professional development varies substantially (Graham, Milanowski, & Miller, 2012), and we do not yet understand what accounts for this variation. Nonetheless, the evidence discussed here implies mitigating observer bias arising from the assignment of observations by state policy is a task worth pursuing.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). The Long Term Impacts of Teachers: Teacher Value Added and Student Outcomes in Adulthood. *American Economic Review*, *104*(9), 2633–2679.

de Barros, A. (2019). Evaluating Teacher Evaluation: Evidence from Chile. *Organization of Schools and Systems & Education in Global Contexts*. Presented at the Society for Research in Educational Effectiveness, Washington, DC.

Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings* (No. 8882021513; pp. 1–33).

Hunter, S. B. (2019a). The Effects of Increasing the Number of Observations Per Teacher on Student Discipline Infractions. *Teacher Evaluation System Design and Teacher Responses*. Presented at the Association for Education Finance and Policy, Kansas City, MO.

Hunter, S. B. (2019b). *The Effects of More Frequent Observations on Student Achievement Scores* (No. 2019–04). Retrieved from Tennessee Education Research Alliance website: https://peabody.vanderbilt.edu/TERA/files/TERA_Working_Paper_2019-04.pdf

Hunter, S. B. (2019c, March). *Unintended Bias in Teacher Observation Scores: The Assignment, Not Receipt, of Classroom Observations*. Presented at the Annual Meeting of the Society for Research on Educational Effectiveness, Washington, D.C.

Jackson, C. K. (2018). What Do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes. *Journal of Political Economy*, *126*(5), 36.

Kraft, M. A., & Gilmour, A. F. (2016). Can Principals Promote Teacher Development as Evaluators? A Case Study of Principals' Views and Experiences. *Educational Administration Quarterly*, *52*(5), 711–753. https://doi.org/10.1177/0013161X16653445

Milanowski, A. (2017). Lower Performance Evaluation Practice Ratings for Teachers of Disadvantaged Students: Bias or Reflection of Reality? *AERA Open*, *3*(1), 233285841668555. https://doi.org/10.1177/2332858416685550

Papay, J. P., & Richard, C. (2017). *Evaluation for Teacher Development: Exploring the Relationship between Features of Teacher Evaluation Systems and Teacher Improvement*. Working Paper, Brown University.

Rigby, J. G. (2015). Principals' Sensemaking and Enactment of Teacher Evaluation. *Journal of Educational Administration*, *53*(3), 374–392. https://doi.org/10.1108/JEA-04-2014-0051

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin Company.

Steinberg, M. P., & Donaldson, M. L. (2016). The New Educational Accountability: Understanding the Landscape of Teacher Evaluation in the Post-NCLB Era. *Education Finance and Policy*, *11*(3). https://doi.org/10.1162/EDFP_a_00186

Steinberg, M. P., & Kraft, M. A. (2017). The Sensitivity of Teacher Performance Ratings to the Design of Teacher Evaluation Systems. *Educational Researcher*, *46*(7), 378–396. https://doi.org/10.3102/0013189X17726752

Tennessee Board of Education. *Teacher and Principal Evaluation Policy*, 5.201 § (2013).

Tennessee Department of Education. (2016). *Evaluation | TEAM-TN*. Retrieved from http://team-tn.org/evaluation/