RedFame

Journal of Education and Training Studies
Vol. 8, No. 4; April 2020
ISSN 2324-805X E-ISSN 2324-8068
Published by Redfame Publishing
URL: http://jets.redfame.com

Examining the Relationship Between Teacher Performance Ratings and District Under the Ohio Teacher Evaluation System

Nicholas P. Elam[1], W. Holmes Finch[1]

[1]Teachers College, Ball State University, Muncie, Indiana, United States

Correspondence: Nicholas P. Elam, Teachers College 909, Ball State University, Muncie, IN 47306, United States.

Received: January 2, 2020 Accepted: February 12, 2020 Online Published: February 13, 2020

doi:10.11114/jets.v8i4.4720 URL: https://doi.org/10.11114/jets.v8i4.4720

Abstract

The soundness of the Ohio Teacher Evaluation System (OTES) depends heavily on evaluators' uniform interpretation of the qualitative Teacher Performance rubric. This study investigates the relationship between teachers' district of employment, and the Teacher Performance ratings they receive under OTES. For Ohio districts that implemented OTES in 2012-2013, 2013-2014, and 2014-2015, the proportion of various Teacher Performance ratings and Student Growth Measures ratings are examined and compared to statewide proportions, using descriptive data and a log-linear model. Findings speak to the importance of a continued or renewed emphasis on fostering uniform interpretation and implementation of teacher evaluation rubrics and systems.

Keywords: teacher evaluation, teacher performance, interrater reliability

1. Introduction

Various stakeholders debate the fundamental purpose of teacher evaluation in K-12 schools. Many regard teacher evaluations as a summative means to assess individual teacher performance and ultimately dismiss underperforming teachers. Many others regard teacher evaluation as a formative means to identify individual strengths and weaknesses and play a valuable role in improving teacher performance going forward. Still others view teacher evaluation as a toothless means to maintain an appearance of accountability for schools and teachers, fueled by a longstanding trend where over 99 percent of teachers receive satisfactory ratings on their evaluations (Harris, 2011; Author).

Over the last generation, states have changed their teacher evaluation systems in various ways in order to increase accountability for individual teachers and for schools, including by: incorporating student achievement data more prominently in the derivation of teacher evaluation ratings, shifting from raw achievement data toward value-added measures, increasing the number of possible ratings classifications, increasing the frequency of evaluations, and attaching more tangible positive and negative consequences to teacher evaluation ratings (Hull, 2013; Doherty & Jacobs, 2015). These new and revised evaluation systems replaced systems that were based only on qualitative observation feedback and/or incorporated crude achievement data, offered only two (satisfactory/unsatisfactory) ratings classifications, called for teachers to be evaluated on a relatively infrequent basis, and rarely led to tangible consequences for teachers.

The state of Ohio introduced the Ohio Teacher Evaluation System (OTES) in 2012-2013. The Ohio Department of Education allowed districts to grandfather in their existing teacher evaluation system until their current collective bargaining agreement expired. OTES resembled many of the new and revised teacher evaluation systems in other states, in that OTES incorporated student achievement data (including value-added data) to a greater extent, included an increased number of ratings classifications, called for teachers to be evaluated more frequently, and carried more tangible consequences for teachers based on their evaluation ratings.

Specifically, OTES derives teacher evaluation ratings 50% from qualitative feedback (Teacher Performance ratings, which include the following classifications: Accomplished, Skilled, Developing, Ineffective) and 50% from quantitative achievement data (Student Growth Measures, which originally included only three classifications, and now include the following five classifications: Most Effective, Above Average, Average, Approaching Average, Least Effective). Teachers' Student Growth Measures ratings are derived differently based on the grade level and subject area they teach: Category A1 teachers teach exclusively in a grade level and subject area where value-added data are available; Category A2 teachers teach part of the time, but not exclusively, in such grade levels/subject areas; Category B teachers teach

other grade levels/subject areas where a vendor assessment is available; and Category C teachers teach other grade levels/subject areas where, in many cases, they must create their own assessments (Ohio Department of Education, 2014).

Teachers ultimately receive one of four overall ratings: Accomplished, Skilled, Developing, or Ineffective. Originally, OTES called for every teacher to be evaluated every year. Beginning in 2014-2015, teachers who receive an Accomplished overall rating may be evaluated once every three years, and teachers who receive a Skilled overall rating may be evaluated once every two years. OTES also uses a misaligned chronological format, where a teacher's Student Growth Measures rating for a given year is derived either from value-added data from the previous year, and/or from non-value-added data from the given year.

In an effort to increase teacher accountability and improve teacher practice, the designers of various new and revised teacher evaluation systems also seek to promote educational equity for students (Tyack & Cuban, 1995; Hess, 1999). However, certain design aspects raise concerns about the equity of the teacher evaluation systems themselves. Regarding OTES, its chronological misalignment and policy change related to frequency of evaluation, might seem to raise concerns. After all, for some – and only some - teachers, evaluators might be influenced by having access to their Student Growth Measures ratings well in advance of submitting a Teacher Performance rating. And for some teachers, evaluators might be tempted to inflate Teacher Performance ratings – pushing them toward an Accomplished or Skilled overall rating – as a way to enjoy a relaxed frequency-of-evaluation guideline for those teachers and to lighten their own evaluation workload going forward.

However, quantitative evidence does not indicate that evaluators are influenced by knowing a teacher's Student Growth Measures rating in advance of submitting a Teacher Performance rating and does not indicate that evaluators inflate ratings as a way to lighten their evaluation workload going forward. Furthermore, qualitative data also do not indicate that OTES evaluators themselves see either of these policy aspects as a reason for concern (Author).

Instead, OTES evaluators are concerned about two other factors – one fundamental to nearly any widespread evaluation system, and one more specific to OTES. OTES evaluators shared a concern that, in a system so dependent on a universal qualitative rubric, OTES could not realistically ensure uniform interpretation and implementation. Also, OTES evaluators expressed great concern about the fairness of a system that calls for Student Growth Measures to be derived from such widely varying sources for different teachers (Author). This study explores each of those concerns.

1.1 Research Questions

- *What relationship exists among Teacher Performance Ratings, Student Growth Measures Ratings, and district conducting the evaluation under the Ohio Teacher Evaluation System?*
- *How does the distribution of evaluation ratings in individual districts compare to the distribution of evaluation ratings in the state of Ohio overall?*
- *How does the distribution of evaluation ratings for teachers subject to a standardized test compare to the distribution of evaluation ratings for teachers subject to a self-created test?*

2. Literature Review

2.1 Interrater Reliability

OTES (and teacher evaluation systems in many other states) is designed for evaluators to use and interpret a Teacher Performance rubric uniformly across the state. The rubric serves as a tool for evaluators to assess teachers within ten different standards (grouped within the broader categories Instructional Planning, Instruction & Assessment, and Professionalism), with each standard including one or more indicators that place teachers as Accomplished, Skilled, Developing, or Ineffective (Ohio Department of Education, 2018).

With such a design, OTES' viability depends greatly on interrater reliability (arguably more so than previous district-based teacher evaluation systems, where districts designed/adopted, weighed, and interpreted their own performance criteria). Previous studies illuminate interrater reliability concerns inherent in teacher evaluation systems. Some literature is more empathetic to principals and evaluators, while other literature is more empathetic to teachers, but much of the literature express a common theme of concern regarding interrater reliability.

OTES, like many other recently-adopted teacher evaluation systems, requires more frequent and intensive observations and documentation from principals and evaluators. On its face, this would appear to promote interrater reliability. However, if the associated increase in time commitment is too steep, principals and evaluators might consciously or unconsciously rush through observations, detrimentally affecting interrater reliability and more generally diminishing the fundamental worthiness of the evaluation process and results (Cosner, Kimball, Barkowski, Carl, & Jones, 2015; Neumerski et al., 2018; Derrington & Martinez, 2019).

Principals face other challenges when conducting evaluations, including making the fine distinction between evaluating *teaching* and evaluating *teachers* (McGreal, 1982). Principals also must consider (for better or worse) their working relationship with a teacher going forward (Neumerski et al., 2018; Derrington & Martinez, 2019), balancing brutal retrospective honesty of what has been observed and encouragement and belief in a teacher's ability going forward (Cosner et al., 2015) – something that peer evaluators do not have to consider in the same way or to the same extent (Manzeske, Eno, Stonehill, Cumming, & MacGillivary, 2014).

Further muddying the issue of interrater reliability, no ideal threshold of consistency exists (Manzeske et al., 2014). And so, while the Ohio Department of Education offers extensive initial and refresher training and credentialing to OTES evaluators, it might be difficult to ever reach a consensus on how much training is necessary, and to what end. Consensus does not exist currently (Ruffini, Makkonen, Tejwani, & Diaz, 2014). Principals tend to believe that the current level of training is sufficient, while teachers believe more training is necessary for evaluators.

Teachers certainly have a vested interest in the competency and integrity of their evaluators – more so now than ever before, with tangible positive and negative consequences linked to their evaluation ratings (Herlihy et al., 2014). Other studies, less empathetic to principals and evaluators, give voice to teachers' concerns (Ruffini et al., 2014).

Various studies point to various reasons teachers should be concerned about interrater reliability and the legitimacy of their evaluation ratings, including natural differences in evaluators' interpretation of rubrics (Chaplin, Gill, Thompkins, & Miller, 2014), evaluators who are influenced more by preconceived notions of a teacher's ability than by what they actually observe (Sergiovanni, Starratt, & Cho, 2014; Whitehurst, Chingos, & Lindquist, 2015), evaluators who appear not to pay attention during formal observations (Shakman, Riordan, Sanchez, Cook, Fournier, & Brett, 2012), evaluators who are generally erratic with their ratings (Sporte, Jiang, & Luppescu, 2014), and evaluators who are consciously or unconsciously biased against teachers with students of low income and/or racial minorities (Chaplin et al., 2014; Whitehurst et al., 2015).

2.2 Necessity and Viability of New Teacher Evaluation Systems

Other studies address a variety of issues that directly or indirectly reaffirm or call into question the necessity and viability of new teacher evaluation systems, many of which are founded on increased incorporation of student achievement data and/or more comprehensive observation rubrics. Some systems are so laborious that even the most highly-rated teachers believe the process to have a negative effect on their job satisfaction and commitment to the profession (Ford, Van Sickle, Clark, Fazio-Brunson, & Schween, 2017).

A number of studies have shown that previous teacher evaluation systems were vulnerable to ratings inflation, with nearly all teachers receiving the highest possible rating (Forman & Markson, 2015). This raises a number of concerns, including a lack of accountability for teachers, and the inability to distinguish teachers of varying quality (Headden & Silva, 2011; Shakman et al., 2012). This phenomenon has contributed greatly to the evolution of teacher evaluation systems, specifically necessitating increased incorporation of objective student achievement data and the attempt to make observation ratings more meaningful and objective through the creation of more nuanced rubrics.

Of course, not just any student achievement data and observation rubrics will do. While student achievement data might appear completely objective in theory, it can be greatly influenced by factors outside of the control of students, teachers, and school leaders (Sergiovanni et al., 2014). With this in mind, many states (including Ohio) incorporate value-added measures as a way to account for these effects. However, many value-added models are flawed, and are inappropriately assumed to be more sound than they truly are (Amrein-Beardsley & Holloway, 2019).

Regardless of whether the value-added model is truly sound, in Ohio, not all teachers have value-added data tied to their particular grade level or subject area, and so Student Growth measures under OTES are derived from widely-varying data sources for various teachers, including many teachers who may create and administer their own achievement tests (Lacireno-Paquet, Morgan, & Mello, 2014). This disparity is the source of one of the strongest and most common concerns by teachers and evaluators related to the design of OTES (Ruffini et al., 2014; Author).

The design of an observation rubric is important, too. Darling-Hammond (2013) found that standard-based rubrics can be conducive to valuable feedback for teachers. Von Frank (2011) found that some – but not too much – granularity can be beneficial in observation rubrics.

This study examines the extent to which a teacher's district influences his or her Teacher Performance rating under OTES. This study speaks in part to the interrater reliability in OTES and speaks more broadly to the overall viability of OTES, an evaluation system designed and heavily reliant on uniform interpretation and implementation of evaluation criteria and materials.

3. Methods

3.1 Participants

Only districts that the Ohio Department of Education identified as having implemented OTES beginning in the earliest possible year (2012-2013) and continued to implement OTES during the 2013-2014 and 2014-2015 school year were considered. (Data from subsequent years were not considered, as the Ohio Department of Education introduced safe harbor provisions to coincide with the administration of new standardized, value-added tests throughout the state. These safe harbor provisions allowed teachers whose Student Growth Measures rating would normally be derived from value-added data, to be derived from self-created tests, and could ultimately influence teachers' evaluation ratings during those years.) Data were obtained via public records request. Among those districts, only those 15 districts whose OTES data were disaggregated by individual teacher were included in the sample.

Overall, data included 2953 individual teacher ratings from 2012-2013, 2013-2014, and 2014-2015. The highest participating district had 403 individual ratings in the sample, whereas the lowest participating district contributed 69 individual ratings. The distributions of individuals by Student Growth Measure and Teacher Performance ratings appear in Table 1.

Table 1. Distribution of ratings by student growth measure category and teacher performance rating

Student Growth Category	Frequency	Percent
Above	927	31.4%
Expected	1745	59.1%
Below	281	9.5%
Teacher Performance Rating		
Accomplished	873	29.6%
Skilled	1992	67.5%
Developing	84	2.8%
Ineffective	4	0.1%

The majority of individual teachers met the expected level of student growth, with an additional 31.4% of teachers whose students demonstrated growth greater than what was expected. Over 97% of the responding teachers were rated as either Skilled or Accomplished.

4. Statistical Analysis

4.1 What Relationship Exists Among Teacher Performance Ratings, Student Growth Measures Ratings, and District Conducting the Evaluation Under the Ohio Teacher Evaluation System?

In order to address the research question regarding the relationships among district, Teacher Performance rating, and Student Growth measures rating categorization, a log-linear model was used. The log-linear model provides estimates of the relationships among categorical variables, such as those of interest here. Two-way and three-way interactions among all combinations of the variables were tested. Statistically significant interactions identified using the log-linear model were followed up with cross-tabulations of the relevant variables, along with standardized cell residuals. Cell residuals are the difference between the observed number of individuals in a given combination of categories (e.g., Student Growth rating above expectations and Teacher Performance rating of skilled) and the number that would be expected if the two variables were not related to one another. These residuals are then standardized so that they can be interpreted as standard normal values, or *Z*-scores. By convention, cells with absolute value standardized residuals greater than 2 are taken to be deviating significantly from what would be expected if the two categorical variables are independent of one another, and thus warrant a close examination (Agresti, 2013). Data analyses were carried out using SAS version 9.4 (SAS Institute, 2017), with maximum likelihood used for estimating the log-linear model parameter estimates.

4.2 To What Extent Do Teacher Performance Ratings in Individual Districts Differ from Statewide Teacher Performance Ratings under the Ohio Teacher Evaluation System?

For each Ohio district with available data that implemented OTES in 2012-2013, 2013-2014, and 2014-2015 (2015-2016 and 2016-2017 data are excluded due to Safe Harbor provisions), the proportion of various Teacher Performance ratings (Accomplished, Skilled, Developing, Ineffective) are examined and compared to statewide proportions (controlling for year and for Student Growth Measures ratings), using descriptive data and the Chi Square Test of Goodness of Fit.

4.3 Do Teachers Receive More Favorable Student Growth Measures Ratings When Using a Self-Created Test?

For all Ohio districts with available data that implemented OTES in 2012-2013, 2013-2014, and 2014-2015 (2015-2016 and 2016-2017 data are excluded due to Safe Harbor provisions), the proportion of various Student Growth Measures ratings (Above Expected Growth, Expected Growth, Below Expected Growth) are examined for teachers subject to a standardized test, and compared to proportions for teachers subject to a self-created test (controlling for year and for Teacher Performance ratings), using descriptive data and the Chi Square Test of Goodness of Fit.

5. Results

5.1 What relationship exists among Teacher Performance Ratings, Student Growth Measures Ratings, and District Conducting the Evaluation Under the Ohio Teacher Evaluation System?

The results of the log-linear analysis appear in Table 2. The interactions between district and Student Growth category, district and Teacher Performance rating, and Student Growth category and Teacher Performance rating were all statistically significant. These results mean that there was a statistically significant relationship between each of these variable pairs; i.e., district with Student Growth category, district with Teacher Performance, and Student Growth with Teacher Performance. There was not a statistically significant 3-way interaction. Given the goals of the study, these results indicate that Student Growth category was related to the district in which the respondent worked, that Teacher Performance rating was also related to the district, and that Teacher Performance rating was related to the Student Growth category.

Table 2. Log-linear model test results

Source	DF	Chi-Square	*p*
District	14	96.43	<.0001
Student Growth	2	5.82	0.0546
District X Student Growth	28	186.73	<.0001
Teacher Performance Rating	3	25.08	<.0001
District X Teacher Performance Rating	28	97.36	<.0001
Student Growth X Teacher Performance Rating	5	15.84	0.0073
District X Student Growth X Teacher Performance Rating	36	39.59	0.3127

As noted above, in order to investigate the nature of the relationships identified by the log-linear model, cross-tabulations and standardized residuals were used. Results for Teacher Performance rating by Student Growth appear in Table 3.

Table 3. Cross-tabulation of teacher performance rating by student growth category: frequency and standardized residual

			Accomplished	Developing	Ineffective	Skilled
Student Growth Category	Above	Frequency	364	24	1	538
		Standardized Residual	7.8*	-.6	-.3	-7.4*
	Below	Frequency	50	11	1	219
		Standardized Residual	-4.5*	1.1	1.1	3.9*
	Expected	Frequency	459	49	2	1235
		Standardized Residual	-4.7*	-.1	-.4	4.6*

Note. *Standardized residuals with an absolute value greater than 2.

Recall that standardized residuals with an absolute value of 2 or more indicate a significant deviation between the observed cell frequency and what would be expected if the two variables are independent of one another. The standardized residuals in Table 3 indicate that the frequency of respondents with a combination of an Accomplished performance rating and Student Growth Above what is expected was greater than would be expected with independence. Likewise, the combinations of a Skilled teacher rating with Below, and with Expected Student Growth also occurred more frequently than would be expected under independence. In contrast, the combination of Skilled Teacher Performance with Student Growth in the Above category, and an Accomplished Teacher Performance with school Student Growth in the Expected or Below categories occurred less frequently than would be expected were the two variables independent of one another.

Based on these results, it can be concluded that respondents who received Accomplished Teacher Performance ratings from evaluators were more likely than would be expected by chance to have students who experienced growth Above what would be expected. In addition, those who received a Skilled Teacher Performance rating from evaluators were more likely than expected by chance to have students with Expected or Below expected levels of growth. In contrast, individuals who received an Accomplished Teacher Performance rating from evaluators were less likely to have students whose growth was at the Expected or Below expected levels.

5.2 To What Extent Do Teacher Performance Ratings in Individual Districts Differ from Statewide Teacher Performance Ratings under the Ohio Teacher Evaluation System?

When examining individual districts, the Chi Square Test of Goodness of Fit did not yield any significant findings. However, select findings are noteworthy using descriptive data

5.2.1 2012-2013

- District 13 gave far fewer Accomplished Teacher Performance ratings than statewide average to teachers with Expected Growth
- District 14 gave no Accomplished Teacher Performance ratings (including teachers above expected growth)
- District 15 gave Skilled rating to 100% of teachers (including teachers Below Expected Growth)
- District 19 Teacher Performance ratings correlated negatively with SGM ratings

5.2.2 2013-2014

- District 3 & 17 Teacher Performance ratings correlated NEGATIVELY with SGM ratings
- District 3 gave many more Accomplished Teacher Performance ratings than statewide average to teachers with Expected Growth
- District 6 gave more Accomplished ratings than statewide average to teachers Above Expected Growth
- District 18 gave far fewer Accomplished ratings to teachers Above Expected Growth
- District 14 & 16 gave fewer Accomplished ratings than statewide average to teachers with Expected Growth
- District 15 gave Skilled rating to 100% of teachers

5.2.3 2014-2015

- District 3 & 9 gave more Accomplished ratings than statewide average to Approaching Average/Average/Above Average teachers
- District 8, 12, 15, & 19 gave fewer Accomplished ratings than statewide average to Approaching Average/Average/Above Average teachers
- District 3 & 18 gave more Accomplished ratings than statewide average to Most Effective teachers
- District 15 & 19 gave fewer Accomplished ratings than statewide average to Most Effective teachers
- District 12 gave Ineffective ratings to two teachers (one Above Average Growth, one Most Effective)
- District 14 gave Skilled rating to 100% of teachers

5.3 Do Teachers Receive More Favorable Student Growth Measures Ratings When Using a Self-Created Test?

Very few districts reported teacher category along with teacher evaluation ratings, and so the Chi Square Test of Goodness of Fit did not yield any significant findings. For the few districts/teachers reported throughout the state, descriptive data from 2013-2014 proves interesting.

Table 4. Distribution of teacher performance ratings by category and student growth measure rating, in 2013-2014

Category A	Frequency	Percent
Above	8	44.4%
Expected	10	55.6%
Below	0	0.0%

Category C	Frequency	Percent
Above	16	88.9%
Expected	2	11.1%
Below	0	0.0%

6. Conclusions/Discussion

This study investigates some of the primary concerns expressed by school principals about the Ohio Teacher Evaluation System. Specifically, this study addresses their concern that a statewide evaluation system so dependent on uniform interpretation would be vulnerable to districts and evaluators who rate teachers more or less favorably than other districts and evaluators, and that teachers whose Student Growth Measures rating is subject to a self-created test are likely to earn more favorable overall evaluation ratings than teachers whose Student Growth Measures rating is subject to a standardized test.

Findings of this study do not support these concerns to a statistically significant extent, even if some select cases support the concerns. Although the two-way interaction between district and Teacher Performance ratings was found to be statistically significant, the current data cannot account for the possibility that some districts might have teachers whose performance truly warrants a higher or lower rating. Student Growth measures is not a perfect way to assess teacher effectiveness, but it does offer a way to account for the varying effectiveness of teachers from one district to the other. Furthermore, the fact that the relationship between teacher rating and student performance was statistically significant, and that the higher rated teachers also had students with higher performance, suggest the possibility that the ratings do reflect actual teacher performance. More work is needed, however, before such a conclusion could be reached definitively. Finally, given that the three-way interaction of district, Teacher Performance rating, *and* Student Growth measures rating is not significant, these results do not indicate that the relationship between teacher rating and student performance varies across districts. In other words, the results indicating that higher rated teachers had students who experienced greater growth were consistent across districts.

Principals' other primary concern – that the design of OTES itself (not, in this case, OTES evaluators) favors some teachers by deriving their Student Growth measures rating from self-created tests, and disfavors other teachers by deriving their Student Growth measures rating from a standardized, value-added test, remains an important topic for further study.

Though the three-way interaction speaks in a way to the merit of OTES, this study raises concerns about OTES in other ways. OTES was designed to be different from previous evaluation systems in a number of ways, including:

- Replacing commonly-used two-tiered evaluation systems with a more refined four-tiered system
- Replacing evaluation systems based primarily/solely on observation data with a system where student achievement data – specifically value-added data – would serve as an integral component of a teacher's overall rating

Given that 97% of teachers were given either Accomplished or Skilled Teacher Performance ratings, and that only 3% of teachers were given either Developing or Ineffective ratings, OTES has, in this way, essentially reverted back to a two-tiered system.

Furthermore, the Ohio Department of Education began offering safe harbor provisions to teachers in 2015-2016 to coincide with the administration of new standardized tests throughout the state. These provisions allowed teachers in value-added grades/subject areas to be assessed in a different way. The rationale for these safe harbor provisions is understandable, but in this way, OTES also reverted back to the evaluation systems that preceded it, by excluding the very data – value-added data – that was meant to distinguish OTES from previous ways of assessing the effectiveness of teachers and schools. The effect these provisions have had on evaluation ratings remains an important topic for further study.

References

Agresti, A. (2013). *Categorical Data Analysis*. Hoboken, NJ: John Wiley & Sons, Inc.

Amrein-Beardsley, A., & Holloway, J. (2019). Value-Added Models for Teacher Evaluation and Accountability: Commonsense Assumptions. *Educational Policy, 33*(3), 516-542. https://doi.org/10.1177/0895904817719519

Chaplin, D., Gill, B., Thompkins, A., Miller, H., & Regional Educational Laboratory Mid-Atlantic. (2014). Professional practice, student surveys, and value-added: Multiple measures of teacher effectiveness in the Pittsburgh Public Schools. REL 2014-024. *Regional Educational Laboratory Mid-Atlantic.*

Cosner, S., Kimball, S. M., Barkowski, E., Carl, B., & Jones, C. (2015). Principal roles, work demands, and supports needed to implement new teacher evaluation. *Mid-Western Educational Researcher, 27*(1), 76-95.

Darling-Hammond, L. (2013). When teachers support and evaluate their peers. *Educational Leadership, 71*(2), 24-29.

Derrington, M. L., & Martinez, J. A. (2019). Exploring Teachers' Evaluation Perceptions: A Snapshot. *NASSP Bulletin, 103*(1), 32-50. https://doi.org/10.1177/0192636519830770

Doherty, K. M. and Jacobs, S. (2015). State of the states 2015: Evaluating teaching, leading, and learning. *National Council on Teacher Quality.* Retrieved from http://www.nctq.org/dmsView/StateofStates2015

Ford, T. G., Van Sickle, M. E., Clark, L. V., Fazio-Brunson, M., & Schween, D. C. (2017). Teacher Self-Efficacy, Professional Commitment, and High-Stakes Teacher Evaluation Policy in Louisiana. *Educational Policy, 31*(2), 202-248. https://doi.org/10.1177/0895904815586855

Forman, K., & Markson, C. (2015). Is "effective" the new "ineffective?" A crisis with the New York state teacher evaluation system. *Journal for Leadership and Instruction, 14*(2), 5-11.

Harris, D. N. (2011). *Value-added measures in education: What every educator needs to know*. Cambridge, MA: Harvard Education Press.

Headden, S., & Silva, E. (2011). Lesson from D.C.'s evaluation system: Teachers give IMPACT low marks on support and professional development. *Journal of Staff Development, 32*(6), 40-44.

Herlihy, C., Karger, E., Pollard, C., Hill, H. C., Kraft, M. A., Williams, M., & Howard, S. (2014). State and Local Efforts to Investigate the Validity and Reliability of Scores From Teacher Evaluation Systems. *Teachers College Record*, *116*(1), 1-28.

Hess, F. M. (1999). *Spinning wheels: The politics of urban school reform.* Washington, DC: The Brookings Institution.

Hull, J. (2013). Trends in teacher evaluation. *Center for Public Education.* http://www.centerforpubliceducation.org/Main-Menu/Evaluating-performance/Trends-in-Teacher-Evaluation-At-A-Glance/Trends-in-Teacher-Evaluation-Full-Report-PDF.pdf

Lacireno-Paquet, N., Morgan, C., & Mello, D. (2014). How states use student learning objectives in teacher evaluation systems: A review of state websites. Regional Educational Laboratory Northeast & Islands.

Manzeske, D. P., Eno, J. P., Stonehill, R. M., Cumming, J. M., MacGillivary, H. L., & Society for Research on Educational Effectiveness. (2014). Assessing teacher effectiveness through dual-rater classroom observations: Researchers and district staff partnering to create calibrated performance evaluations. *Society for Research on Educational Effectiveness.*

McGreal, T. L. (1982). Effective teacher evaluation systems. *Educational Leadership*, *39*(4), 303-05.

Neumerski, C. M., Grissom, J. A., Goldring, E., Rubin, M., Cannata, M., Schuermann, P., & Drake, T. A. (2018). Restructuring Instructional Leadership: How Multiple-Measure Teacher Evaluation Systems Are Redefining the Role of the School Principal. *Elementary School Journal*, *119*(2), 270-297. https://doi.org/10.1086/700597

Ohio Department of Education. (2018). Retrieved from http://education.ohio.gov/Topics/Teaching/Educator-Evaluation-System/Ohio-s-Teacher-Evaluation-System

Ruffini, S. J., Makkonen, R., Tejwani, J., Diaz, M., & Regional Educational Laboratory West. (2014). Principal and teacher perceptions of implementation of multiple-measure teacher evaluation systems in Arizona. REL 2015-062. *Regional Educational Laboratory West.*

SAS Institute, Inc. (2015). *SAS User's Guide*. Cary, NC: SAS Institute, Inc.

Sergiovanni, T. J., Starratt, R. J., & Cho, V. (2014). *Supervision: A redefinition* (9th ed.). New York, NY: McGraw-Hill.

Shakman, K., Riordan, J., Sanchez, M. T., Cook, K. D., Fournier, R., Brett, J., & Regional Educational Laboratory Northeast Islands. (2012). An examination of performance-based teacher evaluation systems in five states: Issues & answers. REL 2012-No. 129. *Regional Educational Laboratory Northeast Islands.*

Sporte, S. E., Jiang, J. Y., Luppescu, S., & Society for Research on Educational Effectiveness. (2014). Teacher evaluation in practice: Understanding evaluator reliability and teacher engagement in Chicago Public Schools. *Society for Research on Educational Effectiveness.*

Tyack, D., & Cuban, L. (1995). *Tinkering toward utopia: A century of public school reform.* Cambridge, MA: Harvard University Press.

Von Frank, V. (2011). Measurement makeover: Florida district revamps teacher evaluation to focus on student achievement. *Journal of Staff Development, 32*(6), 32-39.

Whitehurst, G., Chingos, M. M., & Lindquist, K. (2015). Getting classroom observations right. *Education Next, 15*(1), 62-68.