1-25-2021

Prepping for Another Recession: Re-Assessing the Validity of Teacher Evaluation Systems for Human Capital Decision-Making

Bradley D. Marianno
University of Nevada, Las Vegas, brad.marianno@unlv.edu

Adam Kho
University of Southern California, akho@usc.edu

Tiberio Garza
University of Nevada, Las Vegas, tiberio.garza@unlv.edu

Jonathan Hilpert
jonathan.hilpert@unlv.edu

Prepping for Another Recession: Re-Assessing the Validity of Teacher Evaluation Systems for Human Capital Decision-Making

Bradley D. Marianno, Ph.D.
University of Nevada, Las Vegas

Adam Kho, Ph.D.
University of Southern California

Tiberio Garza, Ph.D.
University of Nevada, Las Vegas

Jonathan Hilpert, Ph.D.
University of Nevada, Las Vegas

Abstract

Problem. The school budget cuts concomitant with the COVID-19 pandemic mean educator jobs may again be threatened by layoffs. During prior recessions, school district administration primarily determined teacher layoffs by virtue of seniority. However, as new evidence emerges that seniority policies may not be the most equitable way to determine teacher layoffs, some have turned towards performance-based measures from evaluation systems. ***Purpose.*** The purpose of this paper is to examine the validity and reliability of the Nevada Educator Performance Framework (NEPF) for making human capital decisions like layoffs. ***Recommendations.*** We recommend that Nevada and other states improve the differentiation in scores across the varying evaluation domains by engaging in more rigorous training of evaluators. Additionally, we recommend that Nevada and other states improve the distribution of final teacher evaluation scores so that the performance measure really distinguishes among teacher performance. Strategies could include lessening the administrative burden of filling out the final evaluation, increasing the number of performance levels, or rotating the specific standards focused on each year.

Introduction

The coronavirus pandemic (COVID-19) has had major consequences for the public education sector. Schools have experienced significant budget cuts resulting in teacher layoffs throughout the nation (Burnette & Will, 2020; Harris & Morton, 2020; Irons, 2020; Turner, 2020), and still more layoffs are expected given decreases in state budgets, reallocations to address other economic and health concerns, and the lack of greater assistance from a federal bailout. Based on similar patterns following the 2008 recession, the layoffs are expected to harm students—particularly Black, Latinx, and low-income students—the most, further widening opportunity and achievement gaps (Jackson, Wigger, & Xiong, 2020). In the case of unavoidable layoffs, making decisions based on teacher effectiveness has shown to harm students less than traditional approaches based on seniority (Boyd et al., 2011; Dabbs, 2020; Goldhaber & Theobald, 2013; Kraft, 2015). However, this requires measures of teacher effectiveness that produce reliable and valid evidence tied to teacher practice and student success.

As a part of the American Recovery and Reinvestment Act of 2009, President Barack Obama launched the Race to the Top federal grant competition, providing grant-based support to states willing to institute educational policies that, in part, overhauled performance evaluation systems for teachers and administrators. States responded with a flurry of legislation aimed at revamping existing evaluation systems. During the 2011 state legislative sessions alone, 19 states enacted comprehensive changes to the way they evaluated teachers and administrators (Marianno, 2015). Over the past decade, almost all states have adopted new teacher evaluation systems (Steinberg & Donaldson, 2016). These policy changes aimed to increase the number of measures used in making determinations of teacher performance, to improve the differentiation in performance between teachers, and to provide decision-makers better information when making difficult layoff, tenure, and dismissal decisions.

In this brief, we first review the literature on trends in educator evaluation systems and prior re-

search that has assessed reliability and validity evidence from these systems. We then turn to the case of Nevada's teacher evaluation system. To support human capital decision-making processes, the Nevada Teachers and Leaders Council created the Nevada Educator Performance Framework (NEPF), first enacted in 2015-16 (Fitzpatrick & Salazar, 2012; Nevada Teachers and Leaders Council, 2013)[1]. Using longitudinal, statewide administrative data, we examine the validity and reliability of the NEPF for making human capital decisions. Our results show that NEPF scores are moderately predictive of student achievement, but we find little distinction in educator domains and little variability in educator ratings that would provide any data for making layoff decisions or other human capital decisions based on teacher effectiveness. We provide recommendations for improving the usefulness of evaluation systems like the NEPF[2].

Recent Trends in Teacher Evaluation Systems

Following the Great Recession of 2007-08, the United States experienced massive educator layoffs (Dabbs, 2020; Felch, Song, & Smith, 2010; Goldhaber et al., 2016; Knight & Strunk, 2016). Traditionally, these layoffs were decided using seniority – "first in, last out" (Boyd et al., 2011; Goldhaber & Theobald, 2013; Sepe & Roza, 2010). However, research emerging from this period began to note the importance of utilizing teacher quality over teacher seniority to make human capital decisions, noting the two were not always highly correlated. While teacher turnover in general harms student achievement (Ronfeldt, Loeb, & Wyckoff, 2013), layoffs made using seniority resulted in greater decreases in student achievement than those made using teacher effectiveness measures, a difference ranging from one-fifth of a standard deviation up to one-third of a standard deviation (Boyd et al., 2011; Dabbs, 2020; Goldhaber & Theobald, 2013; Kraft, 2015). Layoffs based on seniority were also more likely to harm minority students, students from low-income families, and low-performing students, as schools with greater proportions of these student populations are more likely to employ less-experienced teachers (Goldhaber & Theobald, 2013; Knight & Strunk, 2016; Lankford, Loeb, & Wyckoff, 2002; Sepe & Roza, 2010). Further, because teacher salary schedules are based on years of experience, more teacher layoffs would be required under a seniority system to meet budget restraints, which also translates to larger class sizes (Boyd et al., 2011; Kraft, 2015).

In line with this research, an increasing number of states have mandated teacher performance be considered in educator employment decisions, relying on teacher evaluations to provide teacher performance data (Thomsen, 2014). While there is a significant amount of work assessing the predictive validity of individual elements of teacher evaluation systems such as student achievement and student growth measures (Bacher-Hicks, Chin, Kane, & Staiger, 2019; Chetty, Friedman, & Rockoff, 2014; Hill, Kapitula, & Umland, 2011; Kane & Staiger, 2008; Kane et al., 2013; Koedel, Mihaly, & Rockoff, 2015; McCaffrey et al., 2003; Papay, 2011) and classroom observations (Bacher-Hicks et al., 2019; Cohen & Goldhaber, 2016; Garrett & Steinberg, 2015; Goldring et al., 2015; Kane & Staiger, 2012; Kane, Taylor, Tyler, & Wooten, 2011; Steinberg & Garrett, 2016; Whitehurst, Chingos, & Lindquist, 2014), little research has focused on assessing the validity and reliability of the evaluation system as a whole and the specific rating and scoring procedures and scales. In fact, a recent study surveying administrators in a large, suburban school district found administrators were skeptical of the reliability and validity of the evaluation system, yet many states lacked any coherent strategy to assess the reliability and validity of their teacher evaluation systems, despite this concern (Herlihy et al., 2014; Paufler & Clark, 2019).

Examining the Validity of Teacher Evaluation Systems. A small number of studies have published their assessment of educator evaluation systems with a focus on human capital decision-making. Most notably, the New Teacher Project highlighted the Widget Effect, or "the tendency of school districts to assume classroom effectiveness is the same

[1]In addition to providing data to inform human capital decisions, other goals of the NEPF were to foster student learning and growth, improve educators' instructional practices, and engage stakeholders in the process.

[2]Readers can find an extended discussion of our findings and recommendations in our report to the Nevada Legislative Committee on Education at http://crea.sites.unlv.edu/reports/.

from teacher to teacher," treating teachers as interchangeable parts rather than individuals (Weisberg et al., 2009). The study consisted of surveys from 12 districts in four states – Arkansas, Colorado, Illinois, and Ohio. While the districts range greatly in size, location, and management of teachers, each of the 12 districts arrived at the same conclusion. Teacher evaluation systems rarely distinguished effective teachers from ineffective teachers or satisfactory teachers from exceptional teachers. These findings appeared to echo in other states including Florida, Michigan, and Tennessee where 97-98% of teachers were deemed effective (Anderson, 2013). In studies specifically asking principals to assess the performance of teachers, this inability to distinguish effective from ineffective teachers was also pervasive (Jacob & Lefgren, 2008; Lash, Tran, & Huang, 2016).

Related to distinguishing effective from ineffective teachers is the factor structure, or the various aspects of teacher effectiveness assessed by an evaluation system. In most systems, multiple factors are assessed. For instance, the Danielson Framework for Teaching posits four factors in observing teachers and classrooms – Planning and Preparation, Classroom Environment, Delivery of Instruction, and Professional Growth. Each factor is meant to identify a distinct component of teaching effectiveness. However, a study of three large school districts in the southeast and Los Angeles Unified School District found scores only supported a one-factor model, meaning all four proposed factors appeared to measure the same construct (Liu et al., 2019). A similar study evaluating the validity of the National Institute for Excellence in Teaching's (NIET) Teacher Advancement Program (TAP), a widely used observational evaluation framework, also found only one or two factors (depending on method) for a posited three factor structure evaluation system (Sloat, Amrein-Beardsley, & Sabo, 2017).

Lastly, Lash and colleagues (2016) conducted a more comprehensive evaluation of the validity of the Danielson Framework for Teaching classroom observation rubric for Washoe County School District in Nevada. Like prior studies, the evaluation found principals did not identify minimally effective or ineffective teachers, and analysis of the teacher scores indicated a single dimension (or factor) fit the data, though the rubric was designed to measure four different dimensions of teaching. However, teachers' average ratings did show a moderate relationship with student learning, providing some credence to its use as a measure of teaching effectiveness.

Similar to Lash and colleagues (2016), we conduct a more robust validation study of the statewide NEPF. We extend this analysis to include the entire evaluation rating system, including observations and student learning goal measures.

The Nevada Educator Performance Framework. The NEPF is made up of three domains that fall under two overarching categories: educational practice and student outcomes. Educational practice is made up of Instructional Practice and Professional Responsibilities, each with five standards. For standards for each domain, see Appendix A, *Table A1*.

Teachers are rated on a scale of one to four for each domain, and final evaluation ratings are a weighted average of the individual domains on a four-point scale with cutoffs for Highly Effective (3.6 to 4.0), Effective (2.8-3.59), Developing (1.91-2.79), and Ineffective (1.0-1.9). The initial plan for NEPF weighted Instructional Practice 35%, Professional Responsibilities 15%, and student performance 50% of the overall score, where student performance scores were made up of school growth, school proficiency rates, and achievement gap reduction based on the state standardized assessment. However, these weights continued to change annually (with the exception of 2016-17 to 2017-18) in the following years (see Table 1), and in 2016-17, the student performance measure changed from state standardized assessments to a Student Learning Goal (SLG) that provided flexibility for teachers to work with their supervisors to identify student progress goals using assessments other than the state standardized assessment.

In 2014-15, the NEPF was piloted and 125 schools participated in a validation study (WestEd, 2015). Through trainings and telephone interviews with principals, surveys with educators, and focus groups with district superintendents, the study found teachers and administrators believed the framework was valid and reliable. In this study, we utilize administrative data to revisit the reliability and validity of the NEPF five years after initial implementation when the new evaluation system had rolled out and was implemented with all educators in the state and the NEPF was adjusted with new weights to calculate final evaluation scores. Specif-

Table 1. *NEPF Teacher Domain Weights Over Time*

School Year	Domain		
	Instructional Practice	Professional Responsibilities	Student Outcomes
2014-15	35%	15%	50%
2015-16	80%	20%	0%
2016-17	60%	20%	20%
2017-18	60%	20%	20%
2018-19	45%	15%	40%

ically, we ask, can reliable and valid score interpretations be made about teacher effectiveness using data collected from the Nevada Educator Performance Framework? The results of this analysis will be particularly important for understanding the utility of NEPF for human capital decisions as originally designed.

Methods

Data Informing This Brief. The Nevada Department of Education (NDE) provided school-aggregate teacher NEPF scores for the 2015-16 to 2018-19 school years. This data included the number of teachers earning a final rating of ineffective, developing, effective, and highly effective, school average scores on a scale of 1 to 4 for each Instructional Practice and Professional Responsibilities standard, student learning goal scores, and final scores. Individual-level data, including school assignment and grade and subject identifiers, were not included for anonymity purposes. We supplemented this with publicly available Nevada Report Card data, which included school-level student proficiency rates on the annual standardized assessments and school characteristics.

Analytic Strategy. To address whether accurate score interpretations can be made from the NEPF ratings, we examine reliability and validity evidence in a multistep process. We begin by calculating evidence for the internal consistency and dimensionality of NEPF teacher ratings. Then, we calculate aggregate NEPF scores to examine the distribution and score ranges. We conclude with an examination of the predictive validity by fitting an ordinary least squares regression model, predicting student achievement from teacher NEPF scores. More details on our analytic approach can be found in Appendix B and in our full report to the Nevada Legislative Committee on Education (Marianno, Garza, Hilpert, & Kho, 2020).

Results

Internal Consistency and Dimensionality. An internally consistent and valid test is one in which test items that purport to measure the same thing report similar scores across the same respondent. Thinking of the NEPF domains and standards like items on a test, Cronbach's alpha tells us whether a given educator is scoring similarly on the different NEPF standards within a domain. If the standards within a given NEPF domain (say Instructional Practice) are highly correlated with one another (as they should be, if they are truly capturing information on a given teacher's Instructional Practice), then we would expect a high Cronbach's alpha score (above 0.70 on a scale between 0 and 1), and we could conclude that the Instructional Practice domain of the NEPF is internally consistent and reliable. In the case of the Instructional Practice domain, we found a high alpha coefficient of 0.95 with inter-item correlations ranging from 0.65 to 0.80. For the Professional Responsibilities domain, the alpha coefficient was also high at 0.92 with inter-item correlations ranging from 0.62 to 0.83. These results suggest that the NEPF has strong internal consistency.

To establish the tool's validity, it is also useful to explore the dimensionality of the NEPF. Dimensionality has to do with whether the NEPF domains and standards are measuring similar or different things regarding educator performance. By design, the NEPF hypothesizes a two factor structure—it

groups a series of standards under Instructional Practice and a series of standards under Professional Responsibility. We used exploratory factory analysis to examine whether the hypothesized two factor structure consisting of the two NEPF teacher domains of Instructional Practice and Professional Responsibilities best fit the data. Our results suggest that the single factor solution was the best fit to the data. The Instructional Practice and Professional Responsibilities domains load on to the first factor with a correlation of at least 0.76. The results lend support to the idea that the NEPF teacher performance framework is best conceived of as a unidimensional measure of teacher effectiveness – educators scoring highly on the Instructional Practice domain also score highly on the Professional Responsibilities domain.

Distribution of Final Scores. Another indication of validity is whether the NEPF, as a measure of teacher performance, can distinguish between high and low performers. One way to explore this is to look at the amount of variation in the scores. We display summary statistics for the final average scores in Table 2. Given the changes in weighting over the years following implementation of the evaluation system, we do this for unweighted scores as well as for each of the weights from 2017-18 to 2019-20. In all cases, the mean is approximately 3.28, which sits in the middle of the Effective range. In Figure 1, we show the distribution of school-level NEPF teacher final scores. The black vertical lines show the lower and upper bounds of the cut score for a teacher to receive an Effective rating. Without any weighting applied, no schools maintain an average that could be classified as Ineffective (1.9 or lower), and very few maintain an average of Developing. Schools primarily score in the Effective range, with some in the Highly Effective category. These distributions are confirmed in Table 3, which shows final average scores by effectiveness level. Without any weights applied, 92% of schools have a mean score of Effective and another 8% have a mean score of Highly Effective. Less than 1% of schools have a mean score below Effective.

Figure 1. *Distribution of School-Level NEPF Teacher Final Scores (Unweighted)*

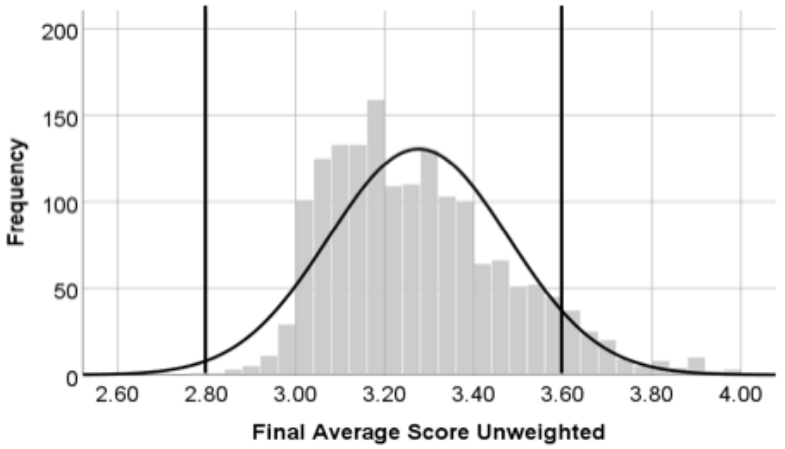

Table 2. *Summary Statistics for School-Level NEPF Teacher Final Scores*

	Mean	SD	Min	Max	Skew	Kurt
Final Avg. Score (Unweighted)	3.28	0.20	2.70	4.00	0.73	0.34
Final Avg. Score (2019-20 weights)	3.27	0.20	2.73	3.99	0.60	0.14
Final Avg. Score (2018-19 weights)	3.28	0.23	2.26	3.99	0.53	-0.03
Final Avg. Score (2017-18 weights)	3.27	0.21	2.68	3.99	0.58	0.09

Note: *Data from all years (2015-16 to 2018-19) are included.*

Predictive Validity on Student Achievement. Lastly, we examine the predictive validity of teacher NEPF scores on student achievement. We use an ordinary least squares regression controlling for school characteristics and year, the results of which are summarized in Table 4. We see small positive associations between teacher NEPF final scores and student achievement, where a 1-percentage point increase in teachers rated Effective or Highly Effective is associated with an increase of approximately 0.01 standard deviations in both reading and math. When we substitute the percentage of teachers rated Effective or Highly Effective with the continuous measure of NEPF final scores, we again see positive associations. Specifically, a 1-point increase in the NEPF Final Score is associated with an 0.24 standard deviation increase in reading and an 0.29 standard deviation increase in math. Overall, our results suggest the NEPF scores are moderately predictive of student achievement. However, the teachers' numeric NEPF scores seem to be more predictive than the final effectiveness ratings.

Table 3. *Percentage of School-Level NEPF Teacher Final Scores Classified by Effectiveness Level*

	Ineffective	**Developing**	**Effective**	**Highly Effective**
Final Avg. Score (Unweighted)	0	0.10	92.20	7.70
Final Avg. Score (2019-20 weights)	0	0.40	92.60	7.00
Final Avg. Score (2018-19 weights)	0	0.70	87.30	12.00
Final Avg. Score (2017-18 weights)	0	0.40	91.50	8.10

Note: *Data from all years (2015-16 to 2018-19) are included.*

Table 4. *Percentage of School-Level NEPF Teacher Final Scores Classified by Effectiveness Level*

	Reading		Math	
	(1)	(2)	(3)	(4)
Percent Teachers Rated Effective or Highly Effective	0.01* (0.00)		0.01* (0.00)	
NEPF Final Score Using 2016-2018 Weighting		0.24* (0.11)		0.29* (0.11)
Year Fixed Effect	X	X	X	X
R-squared	0.456	0.486	0.399	0.433
Observations	1,225	1,194	1,224	1,193

Notes: *Standard errors clustered at the school level in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; Dependent variable = standardized scores derived from uncoarsening total school performance levels by subject and year. Teacher evaluation scores are using 2016-17 and 2017-18 weights. Results are robust to weights from 2018-19 and 2019-20. Data from all years (2015-16 to 2018-19) are included. Models control for student demographic characteristics.*

Discussion and Recommendations

In the past decade, many states have revamped their educator evaluation systems to link teacher performance to student achievement and to better distinguish effective from ineffective teachers. These evaluations have already been used to inform human capital decisions. However, as we likely approach the first recession since Race to the Top, the grant schools received for overhauling their evaluation systems, we can expect more layoffs as a result of state and district budget cuts. With new measures of effectiveness, schools and districts may feel equipped to make layoff decisions based on their new educator effectiveness measures. However, the results of this study caution schools in using these measures in a high-stakes way until the systems are adequately assessed for reliable and valid score interpretation.

Based on our analysis, we make two recommendations as Nevada and other states consider how to improve their teacher evaluation systems. First, states should engage in strategies to improve differentiation in scores between domains. The domains should be related, but the rating scores should load more strongly on their respective factors to demonstrate they are being used to evaluate distinct skills associated with good teaching. One of the goals of the evaluation process is to generate feedback that allows educators to assess opportunities for growth and make progress in those areas. The lack of differentiation between domains, however, means educators may lack clarity on where or how to make improvements or be unable to identify areas of strength. Prior research suggests a significant effort for investment in ongoing training can help (Casabianca, Lockwood, & McCaffrey, 2015). By having raters practice standardized scenarios, raters could gain clarity on more difficult or unclear elements of the evaluation protocol, helping them maintain calibration of their scores with the intended ideal, and thereby improve score differentiation between domains (Park, Chen, & Holtzman, 2014).

Second, we encourage states to improve the distribution of evaluation scores. Our examination of the underlying distributions of the NEPF standard ratings for teachers indicated the full range of the evaluation instrument was not being utilized by evaluators. The accumulation of scores within a narrow scoring band creates a ceiling effect that limits the utility of the evaluation system. Without a clear definition of which teachers are indeed Effective and which are not, it is unclear how to truly make human capital decisions based on this instrument. At best, stakeholders are left to interpret what it means to be a lower level of Effective, for example a score of 3, or to be slightly more Effective at a 3.2, making it difficult to assess teacher growth in meaningful ways. Presumably, when raters make greater use of a greater range of ratings, they can provide greater feedback and incentives for teachers to improve their performance and to distinguish them from Ineffective teachers whose performance has not improved. With little variation in scoring, decisions regarding layoffs may default to alternative criteria like seniority, which further harms students and may have equity implications (Boyd et al., 2011; Dabbs, 2020; Goldhaber & Theobald, 2013; Knight & Strunk, 2016; Kraft, 2015).

The lack of variation in educators' evaluation scores is a problem that many states are still tackling (Kraft & Gilmour, 2017), which could be for several reasons. There is a growing body of research suggesting administrators can get bogged down in deciphering standards and logistical aspects of the evaluation process, spending large amounts of time on evaluations that do not affect positive change (Darling-Hammond, 2015; Marsh et al., 2017; Marshall, 2013; Marzano & Toth, 2013). Further, some school districts require greater reporting and evidence requirements for evaluators who score educators at the bottom or top of the distribution as well as intensive amounts of time providing feedback and support for unsatisfactory teachers (Kraft & Gilmour, 2017). The enhanced paperwork burden associated with scoring educators other than Effective leads to strategic behavior and the clustering of educators at the Effective rating.

We recommend rubrics be detailed enough to provide meaningful standards and indicators reflecting quality teaching while at the same time being simple enough to be used effectively by evaluators in the face of competing time demands. One approach might be increasing the number of performance levels to create truly inadequate levels at the bottom of the scoring range that are rarely used. For instance, splitting the Effective category into two different performance levels. Doing so would expand the scale, thereby helping to limit the ceiling effect that presently exists in the system. States

could replace the single summative rating with a focus on the ratings of individual standards. This would emphasize the specific areas where an educator is succeeding and where they might need additional assistance and could potentially eliminate some discomfort with rating teachers Ineffective overall, another reason principals cited for not differentiating effectiveness (Kraft & Gilmour, 2017). While Nevada's current teacher evaluation system may provide little data to inform human capital decisions during the time of COVID-19, the pandemic provides an opportunity for the state to reset and revisit the validity of the NEPF. While states dropped their accountability assessments and provided flexibility for educator evaluations in the 2019-20 school year, we encourage them to extend that flexibility for the 2020-21 school year as operations are still far from "normal." Instead, states can take this natural pause to examine and reflect on the historical use of their evaluation systems, assess its reliability and validity, and make appropriate changes that will yield a more useful evaluation system when schools return to the new normal.

References

Anderson, J. (2013, March 30). Curious grade for teachers: Nearly all pass. *The New York Times.* Retrieved from https://www.nytimes.com/2013/03/31/education/curious-grade-for-teachers-nearly -all-pass.html

Bacher-Hicks, A., Chin, M. J., Kane, T. J., & Staiger, D. O. (2019). An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys. *Economics of Education Review, 73*, 101919.

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2011). Teacher layoffs: An empirical illustration of seniority versus measures of effectiveness. *Education Finance and Policy, 6*(3), 439–454.

Burnette II, D., & Will, M. (2020, July 14). Thousands of educators laid off already due to COVID-19, and more expected. *Education Week.* Retrieved from https://www.edweek.org/leadership/thousands-of-educators-laid-off-already-due-to-covid-19-and-more-expected/2020/07

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

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review, 104*(9), 2593-2632.

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher, 45*(6), 378-387.

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation, 10*(7).

Dabbs, C. M. (2020). Restricting seniority as a factor in public school district layoffs: Analyzing the impact of state legislation on graduation rates. *Economics of Education Review, 74*, 101926.

Darling-Hammond, L. (2015). *Getting teacher evaluation right: What really matters for effectiveness and improvement.* Teachers College Press.

Felch, J., Song, J., & Smith, D. (2010, December 4). *When layoffs come to L.A. schools, performance doesn't count.* The New Teacher Project. Retrieved from https://tntp.org/news-and-press/ view/when-layoffs-come-to-l.a.-schools-performance-doesnt-count

Fitzpatrick, R., & Salazar, P. (2012). *Teachers and Leaders Council: Summary of anticipated final recommendations and implementation considerations.* Retrieved from https://ccea-nv.org/images/stories/pdfs/TLC_Recommendations_Summary_11_14_12_pdf.pdf.

Garrett, R., & Steinberg, M. P. (2015). Examining teacher effectiveness using classroom observation scores: Evidence from the randomization of teachers to students. *Educational Evaluation and Policy Analysis, 37*(2), 224-242.

Goldhaber, D., Strunk, K. O., Brown, N., & Knight, D. S. (2016). Lessons learned from the Great Recession: Layoffs and the RIF-induced teacher shuffle. *Educational Evaluation and Policy Analysis, 38*(3), 517-548.

Goldhaber, D., & Theobald, R. (2013). Managing the teacher workforce in austere times: The determinants and implications of teacher layoffs. *Education Finance and Policy, 8*(4), 494–527.

Goldring, E., Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make room value added: Principals' human capital decisions and the emergence of teacher observation data. *Educational Researcher, 44*(2), 96-104.

Harris, B., & Morton, N. (2020, September 3). Between COVID-19 and layoffs, schools may not have enough teachers to get through the year. *USA Today.* Retrieved from https://www.usatoday.com/story/news/education/2020/09/03/covid-back-school-layoffs-teaching-jobs/5638287002/

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.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal, 48*(3), 794-831.

Irons, M. E. (2020, June 23). More than 2,000 Massachusetts educators have received layoff or nonrenewal notices. *Boston Globe.* Retrieved from https://www.bostonglobe.com/2020/06/23/metro/more-than-2000-massachusetts-educators-have-received-layoff-or-nonrenewal-notices/

Jackson, C. K., Wigger, C., & Xiong, H. (2020). The costs of cutting school spending: Lessons from the Great Recession. *Education Next, 20*(4), 64–72.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26*(1), 101-136.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment.* [Project] Bill & Melinda Gates Foundation.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (No. w14607). National Bureau of Economic Research.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Bill & Melinda Gates Foundation.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources, 46*(3), 587-613.

Knight, D. S., & Strunk, K. O. (2016). Who bears the costs of district funding cuts? Reducing inequality in the distribution of teacher layoffs. *Educational Researcher, 45*(7), 395-406.

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review, 47*, 180-195.

Kraft, M. A. (2015). Teacher layoffs, teacher quality, and student achievement: Evidence from a discretionary layoff policy. *Education Finance and Policy, 10*(4), 467–507.

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.

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis, 24*(1), 37-62.

Lash, A., Tran, L., & Huang, M. (2016). *Examining the Validity of Ratings from a Classroom Observation Instrument for Use in a District's Teacher Evaluation System.* REL 2016-135. Regional Educational Laboratory West.

Liu, S., Bell, C. A., Jones, N. D., & McCaffrey, D. F. (2019). Classroom observation systems in context: A case for the validation of observation systems. *Educational Assessment, Evaluation and Accountability, 31*(1), 61-95.

Marianno, B. D. (2015). Teachers' unions on the defensive?: How recent collective bargaining laws reformed the rights of teachers. *Journal of School Choice, 9*(4), 551-577.

Marianno, B. D., Garza, T., Hilpert, J., & Kho, A. (2020). *The Nevada Educator Performance Framework: Impact and validity final report.* Retrieved from http://crea.sites.unlv.edu/reports/.

Marsh, J. A., Bush-Mecenas, S., Strunk, K. O., Lincove, J. A., & Huguet, A. (2017). Evaluating teachers in the Big Easy: How organizational context shapes policy responses in New Orleans. *Educational Evaluation and Policy Analysis, 39*(4), 539-570.

Marzano, R. J. and Toth, M. D. (2013) *Teacher evaluation that makes a difference: a new model for teacher growth and student achievement.* ASCD Alexandria, VA.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability.* [Monograph] RAND Corporation.

Nevada Teachers and Leaders Council. (2013). *Nevada Educator Performance Framework teacher and administrator evaluation and support models.* [White Paper]

Osborne, J. W., Costello, A. B., & Kellow, J. T. (2014). *Best practices in exploratory factor analysis* (pp. 86-99). Louisville, KY: CreateSpace Independent Publishing Platform.

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*(1), 163-193.

Park, Y. S., Chen, J., & Holtzman, S. (2014). Evaluating efforts to minimize rater bias in scoring classroom observations. In K. Kerr, R. Pianta, & T. Kane (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching Project* (pp. 383–414). San Francisco, CA: Jossey-Bass.

Paufler, N. A., & Clark, C. (2019). Reframing conversations about teacher quality: school and district administrators' perceptions of the validity, reliability, and justifiability of a new teacher evaluation system. *Educational Assessment, Evaluation and Accountability, 31*(1), 33-60.

Peterson, R. A., & Kim, Y. (2013). On the relationship between coefficient alpha and composite reliability. *Journal of Applied Psychology, 98*(1), 194-198.

Reardon, S. F., Kalogrides, D., & Ho, A. D. (2017). *Linking U.S. school district test score distributions to a common scale.* CEPA Working Paper No. 16-09. Stanford Center for Education Policy Analysis.

Reardon, S., Shear, B., Castellano, K., & Ho, A. (2016). *Using heteroskedastic ordered probit models to recover moments of continuous test score distributions from coarsened data.* CEPA Working Paper No. 16-02. Stanford Center for Education Policy Analysis.

Ronfeldt, M., Loeb, S., & Wyckoff, J. (2013). How teacher turnover harms student achievement. *American Educational Research Journal, 50*(1), 4-36.

Sepe, C., & Roza, M. (2010). *Schools in crisis: Making ends meet. The disproportionate impact of seniority-based layoffs on poor, minority students.* Center on Reinventing Public Education.

Shear, B. R., & Reardon, S. F. (2019). *HETOP: Stata module for estimating heteroskedastic ordered probit models with ordered frequency data.* Version 3.0.

Sloat, E., Amrein-Beardsley, A., & Sabo, K. E. (2017). Examining the factor structure underlying the TAP System for Teacher and Student Advancement. *AERA Open, 3*(4), 2332858417735526.

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), 340-359.

Steinberg, M. P., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure?. *Educational Evaluation and Policy Analysis, 38*(2), 293-317.

Thomsen, J. (2014). *Teacher performance plays growing role in employment decisions. Teacher tenure: Trends in state laws.* Education Commission of the States.

Turner, C. (2020, May 6). *A looming financial meltdown for America's schools.* NPR. Retrieved from https://www.npr.org/2020/05/26/858257200/the-pandemic-is-driving-americas-schools-toward-a-financial-meltdown

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

WestEd. (2015). *A study of the Nevada Educator Performance Framework (NEPF): Year two final report.*

Whitehurst, G., Chingos, M. M., & Lindquist, K. M. (2014). *Evaluating teachers with classroom observations.* Brown Center on Education Policy: Brookings Institute.

Appendix A

Table A1. *NEPF Teacher Standards*

Domain: Instructional Practice	
Standard 1.	New Learning is Connected to Prior Learning and Experience
Standard 2.	Learning Tasks have High Cognitive Demand for Diverse Learners
Standard 3.	Students Engage in Meaning-Making through Discourse and Other Strategies
Standard 4.	Students Engage in Metacognitive Activity to Increase Understanding of and Responsibility for Their Own Learning
Standard 5.	Assessment is Integrated into Instruction
Domain: Professional Responsibilities	
Standard 1.	Commitment to the School Community
Standard 2.	Reflection on Professional Growth and Practice
Standard 3.	Professional Obligations
Standard 4.	Family Engagement
Standard 5.	Student Perception
Domain: Student Outcomes	

Note: *The Student Outcomes domain does not have specific standards—Each is made up of three to four more-specific indicators.*

Appendix B

Below we provide a technical summary of our four step analytic process.

(1) We calculated Cronbach's alpha for the Instructional Practice and Professional Responsibilities domains, estimating the average inter-item correlation among the domain standards (Peterson & Kim, 2013) to examine the internal consistency of NEPF ratings. Then, we use exploratory factor analysis with a promax rotation (Costello & Osborne, 2005) to assess the dimensionality of the NEPF. We hypothesized a two factor structure composed of the standard ratings for the Instructional Practice and Professional Responsibility dimensions. For the NEPF to have adequate dimensionality, the Instructional Practice standard ratings, and the Professional Responsibility standard ratings, respectively, should share more common variance within standards for their respective factors, and less between. To determine the number of factors to retain, we assessed eigenvalues, the scree plot, and item loadings from the pattern matrix, where item loadings for respective factors greater than 0.4 were considered acceptable (Costello & Osborne; 2005; Osborne, Costello, & Kellow, 2014).

(2) We utilize the school-aggregate teacher NEPF scores to explore the domain score ranges and distribution of educator performance on each NEPF domain and standard across all years. Ideally, each NEPF domain and its respective standards should show substantial variation and scoring then follows an approximate normal distribution. In addition to showing the distributions, we present the minimum and maximum scores, standard deviations, skew statistics, and kurtosis statistics.

(3) We examine the predictive validity of NEPF scores on student achievement. We use an ordinary least squares regression in a model estimated as:

$$y_{st}=\beta_0+\beta_1 NEPF_{st}+X_{st}\beta_2+\tau_t+e_{st} \qquad (1)$$

where y_{st} is a measure of student achievement for school *s* in year *t*, as measured on the annual Smarter Balanced Assessment Consortium (SBAC). In particular, we utilize a commonly used uncoarsening procedure to translate frequency counts of students scoring in each performance category on the SBAC (Emerging, Approaching, Meets, Exceeds) into standardized scores (Reardon, Kalogrides, & Ho, 2017; Reardon, Shear, Castellano, & Ho, 2016; Shear & Reardon, 2019). $NEPF_{st}$ represents the school percentage of teachers scoring Effective or Highly Effective. β_1 is the parameter of interest and represents the marginal effect of a percentage point increase in the average school NEPF performance on school achievement. In alternate models, we also use the school average NEPF scores on a continuous scale from 1 to 4.

We control for various time-varying school characteristics using X_{st}, a vector that includes the percentage of male students, students of color, students eligible for free or reduced-price meals (a proxy for students' socioeconomic status), English language learner students, and students with an individualized education plan (IEP). τ_t represents a year fixed effect to account for changes in school growth that are common to all schools in Nevada. To account for multiple observations per school (from different school-by-years), we cluster our standard errors at the school level.