REL Southwest Ask-A-REL Response
Effective Teachers and Principals

August 2017

Question 1:

What is known about the observer (that is, his or her accuracy/consistency across raters, and utility of feedback/alignment with other measures)? What is known about reliable and valid observational measures of teacher quality?

Question 2:

How effective are the (observation) trainings across the country at helping educators appropriately use their states' new educator evaluation systems?

Response:

Following an established REL Southwest research protocol, we conducted a search for research reports as well as descriptive study articles on (1) the reliability and validity of observer ratings and observational measures of teacher quality and (2) the effectiveness of observation trainings at helping educators use their states' new educator evaluation systems. The sources included ERIC and other federally funded databases and organizations, research institutions, academic research databases, and general Internet search engines (For details, please see the methods section at the end of this memo.)

We have not evaluated the quality of references and the resources provided in this response and we offer them only for your reference. Also, we searched the references in the response from the most commonly used resources of research, but they are not comprehensive and other relevant references and resources may exist.

Research References

Question 1: What is known about the observer (accuracy/consistency across raters, utility of feedback/alignment with other measures)? What is known about reliable and valid observational measures of teacher quality?

Bell, C. A., Jones, N. D., Lewis, J. M., & Liu, S. (2013). Predicting observer training satisfaction and certification. *Society for Research on Educational Effectiveness.* https://eric.ed.gov/?id=ED564076

From the ERIC abstract: "The last decade produced numerous studies that show that students learn more from high-quality teachers than they do from lower quality teachers. If instruction is to improve through the use of more rigorous teacher evaluation systems, the implementation of these systems must provide consistent and interpretable information about which aspects of teaching practice need improvement and how those improvements can be accomplished. A primary concern for using observation systems in teacher evaluation is the challenge of training observers to score in valid and reliable ways. In this report the authors will present first-year findings from the "Understanding Consequential Assessment Systems for Teachers" (UCAST) study, which investigates how administrators in a large urban school district learn to use a standardized observation protocol. The relevant research questions are as follows: (1) To what extent do administrator characteristics, beliefs, and expectations predict training satisfaction?; (2) To what extent do administrator characteristics, beliefs, and expectations predict certification success?; and (3) What components of the observation protocol are most challenging for observers to certify, and what accounts for these challenges? This project takes place in the Los Angeles Unified School District (LAUSD). This study is designed to investigate observer thinking and performance as it occurs in practice, and our current analyses draw on quantitative performance and perception data that comes from more than 700 administrators trained by the district during year 1 of the study. To improve the quality of instruction in schools, districts across the country are investing great hopes and resources in the implementation of teacher evaluation systems. Because observation offers a direct measure of instruction and it can point to areas for teacher improvement, additional research on how to implement observation protocols at scale is imperative and will be highly useful to school districts and other education stakeholders. Tables are appended."

Bell, C. A., Yi, Q., Jones, N. D., Lewis, J. M., McLeod, M., & Liu, S.(2014). Observer use of standardized observation protocols in consequential observation systems. *Society for Research on Educational Effectiveness*. https://eric.ed.gov/?id=ED562820

From the ERIC abstract: "Evidence from a handful of large-scale studies suggests that although observers can be trained to score reliably using observation protocols, there are concerns related to initial training and calibration activities designed to keep observers scoring accurately over time (e.g., Bell, et al, 2012; BMGF, 2012). Studies offer little insight into how educational practitioners understand and score observation protocols. This lack of clarity on the factors that facilitate and constrain educators' learning and use of observation systems makes it difficult to implement training and quality control processes at scale. In order to effectively train administrators at scale, it is critical to understand how administrators learn to complete two major tasks--learning to create accurate scores and learning to have conversations around those scores that support instructional improvement. This study takes the perspective that scoring observations of classroom interactions is a complex socio-cognitive process that must be understood in order to improve observer training, and ultimately, score quality. This study suggests administrators have a great deal of knowledge they bring to bear on the observation process. They also bring a commitment to improving instruction to their observation work. In other words, they are not blank slates as they go through observer training. That said, certification data suggest observers have much to learn about how to accurately

score lessons according to the protocol. It is unclear whether the knowledge and commitments observers bring is supportive of high quality scores or the improvement of instruction. Future studies should investigate whether the way in which principals use observation protocols results in better, more useful observation scores. Tables are appended."

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement, 75*(2), 311-337. https://eric.ed.gov/?id=EJ1055077

From the ERIC abstract: "Observations and ratings of classroom teaching and interactions collected over time are susceptible to trends in both the quality of instruction and rater behavior. These trends have potential implications for inferences about teaching and for study design. We use scores on the Classroom Assessment Scoring System-Secondary (CLASS-S) protocol from 458 middle school teachers over a 2-year period to study changes over time in (a) the average quality of teaching for the population of teachers, (b) the average severity of the population of raters, and (c) the severity of individual raters. To obtain these estimates and assess them in the context of other factors that contribute to the variability in scores, we develop an augmented G study model that is broadly applicable for modeling sources of variability in classroom observation ratings data collected over time. In our data, we found that trends in teaching quality were small. Rater drift was very large during raters' initial days of observation and persisted throughout nearly 2 years of scoring. Raters did not converge to a common level of severity; using our model we estimate that variability among raters actually increases over the course of the study. Variance decompositions based on the model find that trends are a modest source of variance relative to overall rater effects, rater errors on specific lessons, and residual error. The discussion provides possible explanations for trends and rater divergence as well as implications for designs collecting ratings over time."

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement, 73*(5), 757-783. https://eric.ed.gov/?id=EJ1019080

From the ERIC abstract: "Classroom observation of teachers is a significant part of educational measurement; measurements of teacher practice are being used in teacher evaluation systems across the country. This research investigated whether observations made live in the classroom and from video recording of the same lessons yielded similar inferences about teaching. Using scores on the Classroom Assessment Scoring System-Secondary (CLASS-S) from 82 algebra classrooms, we explored the effect of observation mode on inferences about the level or ranking of teaching in a single lesson or in a classroom for a year. We estimated the correlation between scores from the two observation modes and tested for mode differences in the distribution of scores, the sources of variance in scores, and the reliability of scores using generalizability and decision studies for the latter comparisons. Inferences about teaching in a classroom for a year were relatively insensitive to observation mode. However, time trends in the raters'

use of the score scale were significant for two CLASS-S domains, leading to mode differences in the reliability and inferences drawn from individual lessons. Implications for different modes of classroom observation with the CLASS-S are discussed."

Cohen, J. & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher, 45*(6), 378-387. https://eric.ed.gov/?id=EJ1112159

From the ERIC abstract: "Improving teacher evaluation is one of the most pressing but also contested areas of educational policy. Value-added measures have received much of the attention in new evaluation systems, but they can only be used to evaluate a fraction of teachers. Classroom observations are almost universally used to assess teachers, yet their statistical properties have received far less empirical scrutiny, in particular in consequential evaluation systems. In this essay, we highlight some conceptual and empirical challenges that are similar across these different measures of teacher quality. Based on a review of empirical research, we argue that we need much more research focused on observations as performance measures. We conclude by sketching out an agenda for future research in this area."

Denny, J. H., Hallam, R., & Homer, K. (2012). A multi-instrument examination of preschool classroom quality and the relationship between program, classroom, and teacher characteristics. *Early Education and Development, 23*(5), 678-696. https://eric.ed.gov/?id=EJ978327

From the ERIC abstract: "Research Findings: A statewide study of preschool classroom quality was conducted using 3 distinct classroom observation measures in order to inform a statewide quality rating system. Findings suggested that Tennessee preschool classrooms were approaching "good" quality on the Early Childhood Environment Rating Scale-Revised (ECERS-R) and provided a mid-to-high emotional and engaging climate as indicated by the Classroom Assessment Scoring System (CLASS) domains of Emotional Support and Student Engagement. However, classrooms were only minimal on the Early Childhood Environment Rating Scale-Extension and the CLASS Instructional Support domain. Past performance on a state quality rating assessment consistently predicted the current quality of preschool classrooms as assessed by all 3 measures. Lead teachers' education in early childhood and experience were also predictors across quality measures. Practice or Policy: Tennessee preschool classrooms scored higher on the ECERS-R, which is the measure utilized in the statewide Quality Rating and Improvement System. However, classrooms generally performed poorly on measures of instructional support and curriculum. This finding illuminates the importance of the tool selected to measure quality in state quality rating and improvement systems and has implications for policy as states work to build systems that enhance quality in early care and education. (Contains 6 tables.)"

Gill, B., Shoji, M., Coen, T., & Place, K. (2016). The content, predictive power, and potential bias in five widely used teacher observation instruments. REL 2017-191. *Regional Educational Laboratory Mid-Atlantic.* https://eric.ed.gov/?id=ED569941

From the ERIC abstract: "School districts and states across the Regional Educational Laboratory Mid-Atlantic Region and the country as a whole have been modifying their teacher evaluation systems to identify more effective and less effective teachers and provide better feedback to improve instructional practice. The new systems typically include components related to student achievement growth and instruments for observing and rating instructional practice. Many school districts and states are considering adopting commercially available instruments for the instructional practice component of their evaluation systems. Yet little data are available to help districts and states choose among available instruments or determine which dimensions of instructional practice merit the greatest emphasis. Most existing data comparing different observation instruments, including their statistical characteristics and their relationship to student achievement, come from the Bill & Melinda Gates Foundation's Measures of Effective Teaching project (Kane & Staiger, 2012). This study examined data from the Measures of Effective Teaching project to address three research questions that might inform district and state decisions about selecting and using five widely used teacher observation instruments: the Classroom Assessment Scoring System, the Framework for Teaching, the Protocol for Language Arts Teaching Observations, the Mathematical Quality of Instruction, and the UTeach Observational Protocol. Specifically, the research questions focused on the major differences and similarities in the dimensions of instructional practice rated by the five observation instruments, whether some dimensions of instructional practice consistently show stronger correlations with teachers' value-added scores across the different observation instruments, and the extent to which characteristics of students in the classroom affect instrument scores. Key findings include: (1) Eight of ten dimensions of instructional practice are common across all five examined teacher observation instruments, demonstrating that large parts of the various instruments are conceptually consistent; (2) All seven of the dimensions of instructional practice with quantitative data are modestly but significantly related to teachers' value-added scores; (3) The classroom management dimension is most consistently and strongly related to teachers' value-added scores across instruments, subjects, and grades; and (4) The characteristics of students in the classroom affect teacher observation scores for some instruments and subjects. Observation scores for English language arts classes may be more susceptible to classroom composition effects. For two of the three instruments (Framework for Teaching and Classroom Assessment Scoring System) used to score English language arts instruction, teachers with a larger percentage of racial/ethnic minority students in their classroom tend to receive lower observation scores; a similar effect was observed with the Framework for Teaching for teachers with lower-achieving students. There was no evidence that the composition of students in the classroom affects scores for the Protocol for Language Arts Teaching Observations (the third instrument used to score English language arts instruction), and there was little indication that student characteristics affect observation scores in math classes. The following are appended: (1) Detailed study methodology; (2) Imputation methodology for value-added model estimation; and (3) Supplementary results."

Halpin, Peter F. (2016). Latent class models for teacher observation data. *Society for Research on Educational Effectiveness.* https://eric.ed.gov/?id=ED567240

From the ERIC abstract: "Recent research on multiple measures of teaching effectiveness has redefined the role of in-classroom observations in teacher evaluation systems. In particular, most states now mandate that teachers are observed on multiple occasions during the school year, and it is increasingly common that multiple raters are utilized across the different rating occasions. In-classroom observations are typically conducted using a rating rubric. However, related research has found that many rubrics measure multiple dimensions of instructional quality, suggesting that teachers' practices are not well described in terms of a total score. Halpin & Kieffer (2015) argued for the use of latent class analysis (LCA) as a means of capturing the multidimensional features of rating rubrics, while also providing the standard error of measurement for each teacher, and item-level diagnostic information that can be used as the basis of feedback to educators and for professional development. The main purpose of the present research is to develop a multilevel extension of the LCA methodology described by Halpin & Kieffer (2015). For a given rating rubric, the multilevel LCA approach is specifically intended to answer the following questions: (1) How reliably (precisely) is a teacher's teaching ability measured during any single observation session?; (2) How consistently does a teacher perform over observation sessions?; and (3) For a given teacher, how many observation sessions are required before his/her teaching ability has been measured with a desired level of precision? The last question in particular has relevance for policy, in that multi-rater systems can place heavy financial demands on school districts in terms of deploying a sufficient number of trained raters to meet required number of observation sessions per teacher. The proposed methodology allows for decisions about the required number of observations to be made on a teacher-by-teacher basis, and to be informed by the data collected from each teacher. Two figures are appended."

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. https://eric.ed.gov/?id=EJ944358

From the ERIC abstract: "Research continues to find large differences in student achievement gains across teachers' classrooms. The variability in teacher effectiveness raises the stakes on identifying effective teachers and teaching practices. This paper combines data from classroom observations of teaching practices and measures of teachers' ability to improve student achievement as one contribution to these questions. We find that observation measures of teaching effectiveness are substantively related to student achievement growth and that some observed teaching practices predict achievement more than other practices. Our results provide information for both individual teacher development efforts, and the design of teacher evaluation systems. (Contains 15 footnotes and 7 tables.)"

Lazarev, V., Newman, D., & Sharp, A. (2014). Properties of the multiple measures in Arizona's teacher evaluation model. REL 2015-050. *Regional Educational Laboratory West*. https://eric.ed.gov/?id=ED548027

From the ERIC abstract: "This study explored the relationships among the components of the Arizona Department of Education's new teacher evaluation model, with a particular focus on the extent to which ratings from the state model's teacher observation instrument differentiated higher and lower performance. The study used teacher-level evaluation data collected by the Arizona Department of Education from five participating pilot LEAs during the 2012/13 school year. The study relied primarily on descriptive statistics calculated from the results of the different component metrics piloted in these LEAs, as well as analysis of the correlations among these components. Results indicated that teachers' observation item scores tended to concentrate at the Proficient level (the second-highest score on a four-point scale: Unsatisfactory, Basic, Proficient, and Distinguished), with this level accounting for 62 percent of all observational item scores. In addition, while the strength of the correlation between results from observations and the state's student academic progress metric was generally low, the correlation varied significantly between high- and low-performing teachers, as well as between certain teacher subgroups. The following are appended: (1) Arizona Teacher Evaluation Model; (2) Pilot Local Education Agency Information; (3) Study Methods; (4) Detailed Arizona Teacher Evaluation Model Pilot Results; 2012/13; and (5) Detecting nonlinear Relationships between Observation Item Scores and Student Academic Progress Metrics."

Manzeske, D. P., Eno, J. P., Stonehill, R. M., Cumming, J. M., & MacGillivary, H. L. (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.* https://eric.ed.gov/?id=ED562698

*From the ERIC abstract: "*Federal policies (e.g., 2002 reauthorization of the Elementary and Secondary Education Act [ESEA] and the American Recovery and Reinvestment Act) posit that teacher quality is a potential leverage point for improving student achievement (U.S. Department of Education, 2010). Moreover, in the Race to the Top competition, teacher effectiveness must be based, in part, on teacher performance measured by classroom observations. This has driven many districts to adopt teacher classroom observation rubrics to meet the Race to the Top requirement. Without clear guidance on how to rate teachers and without proper calibration activities, scores on these rubrics can become upwardly biased, leading to an inability to distinguish among teachers at different performance levels (see, for example, Weisberg, Sexton, Mulhern, & Keeling, 2009). When a rubric is used inconsistently, teachers may not receive useful feedback, and the rubric could lack teacher buy-in, resulting in views that the evaluation does not provide credible information. In partnership with district central office staff, a study was conducted in a set of 20 elementary and middle schools in the western United States to investigate the use of a classroom observation rubric within the context of a district's pilot teacher performance evaluation system. The district wanted to know whether peer raters would use the observation rubric differently compared with principals, whether interrater reliability differed by rater type, and whether district-selected raters, on average, were more or less lenient than principals in rating other teachers."

McCaffrey, D. F., & Casabianca, J. M. (2014). Effect of observation mode on measures of teaching. *Society for Research on Educational Effectiveness*. https://eric.ed.gov/?id=ED563034

From the ERIC abstract: "As the education reform movement increasingly focuses on teachers and teaching, educators, policy-makers, and researchers need valid and reliable measures that can be used to evaluate individual teachers, provide guidance for improving teaching performance, and support research in ways that advance instruction and classroom dialog and practice. A new generation of classroom evaluation tools has recently been developed to support evaluation of teaching. Live observations tend to be the standard for studies of teaching and teacher evaluations in practice. They have the benefit of the observer being in the teacher's physical classroom. This is valuable for teacher evaluations because it gives observation scores credibility among teachers. Using video provides particular affordances because they create a permanent record and teachers can review them to evaluate their own instruction as professional development (Miller, 2007; Sherin & Han, 2004; van Es & Sherin, 2010). Videos can be scored by multiple raters, which can reduce error by averaging scores. The use of video also allows for scores to be audited as a part of quality control and videos can be evaluated using multiple scoring protocols to assess the robustness of inferences to a protocol. For most of these reasons, many recent studies of classrooms have made use of videos (Bill and Melinda Gates Foundation, 2012). Given these affordances, an important issue is to understand the comparability of the nature and quality of information created through these two observation modes. Nearly 20 years ago Jaeger (1993) identified mode of observation as potentially contributing to the psychometric properties of measuring teaching, but little research on mode effects has occurred since. This study is a first step toward rectifying the dearth in knowledge on the effects of observation mode on the psychometric properties of classroom teaching evaluations. It tests for observation mode effects on inferences about teaching, classrooms, and teachers. Research questions include the following: (1) Do raters systematically give higher scores using one observation mode or the other? (2) Does the observation mode affect the rank ordering of scores? (3) Does the observation mode affect the size of the standard errors of measurement or the reliability of scores? (4) What are the implications of errors for inferences about the teaching in a lesson or for a classroom for a year? The authors use data collected for the study Toward an Understanding of Classroom Context (TUCC) to test for mode effects in the scores and inferences about the teaching in lessons and in classrooms. TUCC took place in middle and high schools in an urban fringe mid-Atlantic school district that serves roughly 90 percent students of color and 55 percent students who are eligible for free or reduced price meals. The study concluded that there are small mode effects on score means but they are relatively small and most likely inconsequential. Modes do not rank order teachers differently; it is the measurement error that results in differences between live observations and video scoring in the ordering of classrooms or lessons. Differences in the decomposition of variance across modes are a result of the differences in scoring dates. Scoring trends and the differences in timing across the modes are the only significant difference between modes. They have important implications for studies using classroom observations. Live scoring will confound rater learning with lessons and video scoring can avoid this confound. Figures are appended."

Reddy, L. A., Fabiano, G. A., & Dudek, C. M. (2013). Concurrent validity of the classroom strategies scale for elementary school--observer form. *Journal of Psychoeducational Assessment, 31*(3), 258-270. https://eric.ed.gov/?id=EJ1011849

From the ERIC abstract: "The present study is an initial investigation of the concurrent validity of a new assessment, the Classroom Strategies Scale (CSS version 2.0) for Elementary School--Observer Form. The CSS assesses teachers' use of instructional and behavioral management strategies. In the present study, the CSS is compared to the Classroom Assessment Scoring System (CLASS), a widely researched measure of global classroom quality. In a sample of 125 general education K-5 grade teachers, correlations were computed to assess the relationship between the CSS scales and conceptually similar and dissimilar domains and dimensions on the CLASS. In comparison to the CLASS, the CSS classroom observations and strategy rating scale scores demonstrated correspondence with conceptually similar scales, providing initial evidence for the concurrent and discriminant validity of the CSS. Results highlight the unique features of the CSS for assessing teacher classroom practices. (Contains 5 tables and 1 note.)"

Reinke, W. M., Stormont, M. H., Keith C., Wachsmuth, S. & Newcomer, L. (2015). The brief classroom interaction observation-revised: An observation system to inform and increase teacher use of universal classroom management practices. *Journal of Positive Behavior Interventions, 17*(3), 159-169. https://eric.ed.gov/?id=EJ1064040

From the ERIC abstract: "Schools are increasingly using multi-tiered prevention models to address the academic and behavior needs of students. The foundation of these models is the implementation of universal, or Tier 1, practices designed to support the academic and behavioral needs of the vast majority of students. To support teachers in the use of effective Tier 1 classroom practices, researchers and practitioners need reliable and valid measures of these practices that are sensitive to change over time. The purpose of this study was to examine the reliability and validity of the "Brief Classroom Interaction Observation-Revised" (BCIO-R), which is a direct observation measure of classroom Tier 1 instructional and classroom management practices for use in elementary school classrooms. Findings indicate that the BCIO-R can be reliably implemented in the classroom context. In addition, the measure is associated with important teacher-reported constructs such as efficacy in classroom management and burnout. Furthermore, the measure is sensitive to change as indicated by demonstration of improvement in classroom management variables among teachers who received a universal classroom management intervention versus teachers who did not receive training. Having reliable and valid measures to evaluate and monitor teacher use of universal classroom practices can be useful when consulting to support teachers and improve student outcomes."

Rusby, J. C., Crowley, R., Sprague, J., & Biglan, A. (2011). Observations of the middle school environment: The context for student behavior beyond the classroom. *Psychology in the Schools, 48*(4), 400-415. https://eric.ed.gov/?id=EJ921364

From the ERIC abstract: "This article describes the use of an observation system to measure middle school staff practices, environment characteristics, and student behavior in the school common areas. Data were collected at baseline from 18 middle schools participating in a randomized controlled trial of school-wide Positive Behavior Support.

The observations were reliable and showed sensitivity to differences between school settings and between schools. Multilevel models with students nested in schools were used to examine the associations of staff practices and the school environment with student behavior. Less effective behavior management and more staff criticism, graffiti, and percentage of low-income students were associated with student problem behaviors. Greater use of effective behavior management and positive attention, and fewer low-income students were associated with positive student behavior. The use of data-based feedback to schools for intervention planning and monitoring is illustrated. Implications for school-wide efforts to improve student behavior in middle schools are discussed. (Contains 6 tables and 1 figure.)"

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. https://eric.ed.gov/?id=EJ1100448

From the ERIC abstract: "As states and districts implement more rigorous teacher evaluation systems, measures of teacher performance are increasingly being used to support instruction and inform retention decisions. Classroom observations take a central role in these systems, accounting for the majority of teacher ratings upon which accountability decisions are based. Using data from the Measures of Effective Teaching study, we explore the extent to which classroom composition influences measured teacher performance based on classroom observation scores. The context in which teachers work--most notably, the incoming academic performance of their students--plays a critical role in determining teachers' measured performance. Furthermore, the intentional sorting of teachers to students has a significant influence on measured performance. Implications for high-stakes teacher accountability policies are discussed."

Question 2: How effective are the (observation) trainings across the country at helping educators appropriately use their states' new educator evaluation systems?

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. https://eric.ed.gov/?id=EJ1051625

From the ERIC abstract: "Policy makers at the federal level have embraced an educator effectiveness agenda, which in turn has driven many states across the country to rapidly develop and implement new and more complex teacher evaluation systems. It is increasingly clear that the success of these nascent teacher evaluation systems partly depends on the will, skill, and capacity of school principals, individuals who have historically been tasked with evaluating teachers. School principals have traditionally had, and will in most cases continue to have, primary responsibility for evaluating the 3.7 million public school teachers nationwide. While teacher evaluation innovations present several opportunities for improving instructional supervision and teacher quality, they also involve several challenges, especially on the part of principals. Time demands and cognitive challenges will be inevitable as principals learn about and implement new teacher evaluation systems. Simultaneously, other educational changes going to scale, including Common Core State Standards with aligned assessments and state school

accountability systems, will compete for the attention of school leaders and teachers. Negotiating these changes to maximize the positive potential of evaluation reforms requires a commitment by states and districts to resources for training and support as well as policy coherence."

Fuller, E. J., Hollingworth, L., & Liu, J. (2015). Evaluating state principal evaluation plans across the United States. *Journal of Research on Leadership Education, 10*(3), 164-192. https://eric.ed.gov/?id=EJ1087211

From the ERIC abstract: "Recent federal legislation has created strong incentives for states to adopt principal evaluation systems, many of which include new measures of principal effectiveness such as estimates of student growth and changes in school climate. Yet, there has been little research on principal evaluation systems and no state-by-state analysis of the principal evaluation systems adopted at the behest of the legislation. This study uses survey data and document review to assess the components of principal evaluation systems in the 50 states and Washington, D.C. Finally, based on recent research, this study critiques the various components of these new evaluation systems."

Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T. A., Goldring, E., & Schuermann, P. (2017). Central office supports for data-driven talent management decisions: Evidence from the implementation of new systems for measuring teacher effectiveness. *Educational Researcher, 46*(1), 21-32. https://eric.ed.gov/?id=EJ1132548

From the ERIC abstract: "School districts increasingly push school leaders to utilize multiple measures of teacher effectiveness, such as observation ratings or value-added scores, in making talent management decisions, including teacher hiring, assignment, support, and retention, but we know little about the local conditions that promote or impede these processes. We investigate the barriers to principals' use of teacher effectiveness measures in eight urban districts and charter management organizations that are investing in new systems for collecting such measures and making them available to school leaders and the supports central offices are building to help principals overcome those barriers. Interviews with more than 175 central and school leaders identify barriers in three main areas related to accessing measures, analyzing them, and taking action based on their analysis. Supports fall into four categories: professional development, connecting principals to sources of expertise, creating new structures or tools, and building a data use culture. Survey analysis suggests that indeed principals in high support systems perceive lower barriers to data use and report greater incorporation of teacher effectiveness measures into their talent management decisions."

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. http://journals.sagepub.com/doi/pdf/10.1177/0013161X16653445 and http://scholar.harvard.edu/files/mkraft/files/principals_as_evalutors_3.5_0.pdf .

From the abstract: "New teacher evaluation systems have expanded the role of principals as instructional leaders, but little is known about principals' ability to promote teacher development through the evaluation process. We conducted a case study of principals'

perspectives on evaluation and their experiences implementing observation and feedback cycles to better understand whether principals feel as though they are able to promote teacher development as evaluators. Research Method: We conducted interviews with a stratified random sample of 24 principals in an urban district that recently implemented major reforms to its teacher evaluation system. We analyzed these interviews by drafting thematic summaries, coding interview transcripts, creating data-analytic matrices, and writing analytic memos. Findings: We found that the evaluation reforms provided a common framework and language that helped facilitate principals' feedback conversations with teachers. However, we also found that tasking principals with primary responsibility for conducting evaluations resulted in a variety of unintended consequences which undercut the quality of evaluation feedback they provided. We analyze five broad solutions to these challenges: strategically targeting evaluations, reducing operational responsibilities, providing principal training, hiring instructional coaches, and developing peer evaluation systems. Implications: The quality of feedback teachers receive through the evaluation process depends critically on the time and training evaluators have to provide individualized and actionable feedback. Districts that task principals with primary responsibility for conducting observation and feedback cycles must attend to the many implementation challenges associated with this approach in order for next-generation evaluation systems to successfully promote teacher development."

Riordan, J., Shakman, K., Chang, Q., Lacireno-Paquet, N., & Bocala, C. (2015). Redesigning teacher evaluations: Lessons from a pilot implementation. Stated Briefly. REL 2016-101. *Regional Educational Laboratory Northeast & Islands.* https://eric.ed.gov/?id=ED561234

From the ERIC abstract: "This "Stated Briefly" report is a companion piece that summarizes the results of another report of the same name. REL Northeast and Islands, in collaboration with the Northeast Educator Effectiveness Research Alliance and the New Hampshire Department of Education conducted a study of the implementation of new teacher evaluation systems in New Hampshire's School Improvement Grant schools. While the basic system features are similar across district plans, the specifics of these features vary considerably by district. Further, district fidelity to the plans, as measured by the exposure of teachers to different features of the evaluation system, ranged from moderate to high. Finally, researchers identified several factors related to implementation: (1) capacity of administrators to conduct evaluations; (2) initial and on-going evaluator training; (3) the introduction and design of student learning objectives; (4) the professional climate of schools, including the support of the new system by teachers and evaluators. [For the full report, see ED552484.]"

Additional Organizations to Consult

Questions 1 and 2:

Center on Great Teachers and Leaders at American Institutes for Research - http://www.gtlcenter.org/content/gtl-overview

From the website: "The Center on Great Teachers and Leaders (GTL Center) is dedicated to supporting state education leaders in their efforts to grow, respect, and retain great teachers and leaders for all students. The GTL Center continues the work of the National Comprehensive Center for Teacher Quality (TQ Center) and expands its focus to provide technical assistance and online resources designed to build systems that

Support the implementation of **college and career standards**.

Ensure the **equitable access** of effective teachers and leaders.

Recruit, retain, reward, and support **effective educators**.

Develop coherent **human capital management systems.**

Create **safe academic environments** that increase student learning through positive behavior management and appropriate discipline.

Use **data** to guide professional development and improve instruction.

Teacher Evaluation and Training Information for the Five REL Southwest States

Arkansas Department of Education—Teacher Excellence And Support System (TESS) http://www.arkansased.gov/public/userfiles/HR_and_Educator_Effectiveness/TESS/Handbook%20Jan%202016.pdf

Louisiana Department of Education—LA Compass Toolbox—http://www.louisianabelieves.com/resources/classroom-support/teacher-support-toolbox/compass-teacher-results

New Mexico Public Education Department—http://ped.state.nm.us/ped/NMTeach_Toolbox.html

Oklahoma State Department of Education—OK Teacher and Leader Effectiveness (TLE) Requirements http://sde.ok.gov/sde/tle

Texas Education Agency—Texas Teacher Evaluation and Support System http://tea.texas.gov/Texas_Educators/Educator_Evaluation_and_Support_System/Texas_Teacher_Evaluation_and_Support_System/

Methods

Keywords and Search Strings

The following keywords and search strings were used to search the reference databases and other sources:

Observation

Observer

Observed rating

Measure

State

Evaluation

Assessment

Training

Reliable

Valid

Teacher quality observation

Professional development observation

Databases and Resources

We searched ERIC for relevant resources. ERIC is a free online library of over 1.6 million citations of education research sponsored by the Institute of Education Sciences. Additionally, we searched Google Scholar and PsychInfo.

Reference Search and Selection Criteria

When we were searching and reviewing resources, we considered the following criteria:

Date of the publication: The most current information (primarily published from 2011 to the present) is included.

Search Priorities of Reference Sources: Search priority is given to study reports, briefs, and other documents that are published and/or reviewed by IES and other federal or federally funded organizations, academic databases, including ERIC, EBSCO databases, JSTOR database, PsychInfo, PsychArticle, and Google Scholar.

Methodology: Following methodological priorities/considerations were given in the review and selection of the references: (a) study types – randomized control trials, quasi experiments, surveys, descriptive data analyses, literature reviews, policy briefs, etc., generally in this order (b) target population, samples (representativeness of the target population, sample size, volunteered or randomly selected, etc.), study duration, etc. (c) limitations, generalizability of the findings and conclusions, etc.

This memorandum is one in a series of quick-turnaround responses to specific questions posed by stakeholders in the Southwest Region (Arkansas, Louisiana, New Mexico, Oklahoma, and Texas), which is served by the Regional Educational Laboratory (REL) Southwest at SEDL. This memorandum was prepared by REL Southwest under a contract with the U.S. Department of Education's Institute of Education Sciences (IES), Contract ED-IES-12-C-0012, administered by SEDL. Its content does not necessarily reflect the views or policies of IES or the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.