CALDER Policymakers Council

Research Brief

ASSESSING THE EVIDENCE ON TEACHER EVALUATION REFORMS

Cara Jackson

Bellwether Education Partners

James Cowan

American Institutes for Research/CALDER

Suggested citation:

Cara, J., & Cowan, J. (2018). *Assessing the Evidence on Teacher Evaluation Reforms* (CALDER Research Brief). Washington, D.C.: National Center for Analysis of Longitudinal Data in Education Research. CALDER Policy Brief No. 13-1218-1.

The crafting and dissemination of this research brief was supported by the National Center for the Analysis of Longitudinal Data in Education Research (CALDER), which is funded by a consortium of foundations. For more information about CALDER funders, see www.caldercenter.org/about-calder. Note that the views expressed are those of the authors and do not necessarily reflect those of our funders or the institutions to which the authors are affiliated.

Assessing the Evidence on Teacher Evaluation Reforms

Cara Jackson

Bellwether Education Partners

James Cowan

American Institutes for Research/CALDER

CALDER Policy Brief No. 13-1218-1

Highlight bullets

- Some new teacher evaluation systems include measures of teacher effectiveness with well-documented connections to student outcomes, although existing measures may fail to capture important teaching skills and, in many locations, final evaluation results still fail to meaningfully differentiate teacher performance.
- Rigorous evaluations of policies that link teacher evaluation with differentiated compensation have shown mixed results, but well-designed policies supported by teacher evaluation and professional development reforms appear to have improved student achievement in several districts.
- Even low-stakes evaluation systems may affect the composition of the teaching profession by increasing retention among high-performing teachers or encouraging attrition of lower performing teachers.
- We have a limited understanding of how and whether better evaluation improves the practice of incumbent teachers, and the field needs more research on promising strategies to integrate teacher evaluation into professional development programs.

Executive summary

In the last decade, school systems have considered a variety of reforms to teacher evaluation systems. The most notable changes have been incorporating new measures of teacher effectiveness and linking evaluation results to compensation or dismissal (McGuinn, 2015; Steinberg & Donaldson, 2016). In this research brief, we consider a few topics where research has reached conclusions about the effects of reforms and others where research suggests important facts are still unknown.

How well do teacher evaluation metrics measure teacher quality?

A large body of research evidence suggests that observational ratings, value-added, and other commonly used evaluation metrics are associated with student learning outcomes. But policymakers should also be mindful of well-documented limitations of these measures. They imprecisely predict future performance, and, in any given year, some teachers will be miscategorized as high or low performers. Most "multiple-measures" evaluation systems address this limitation by combining several sources of data on teacher effectiveness. Including multiple measures is also meant to capture a wider range of teaching skills, although new research questions whether the most common of these capture teachers' effects on students' long-run educational outcomes.

How well do evaluation policies support districts' development objectives?

Improving the practice of existing teachers is a central promise of teacher evaluation systems. There is some evidence that the introduction of low-stakes evaluation policies can improve teacher effectiveness (Garet et al., 2017; Steinberg & Sartain, 2016; Taylor & Tyler, 2012). But although this evidence is promising, we only have evidence from a few selected locations. Moreover, we know little about how individual components of evaluation reforms affect teacher outcomes. One particularly important unexplored issue is how best to integrate performance evaluations into professional development programs so that teacher evaluation best serves both summative and formative purposes.

Have evaluation reforms changed the composition of the teacher workforce?

Teacher evaluation reforms have been shown to affect the composition of the workforce, although there is limited evidence about the importance of program design or the strength of implementation. Some of the more prominent reforms, which combine comprehensive reforms to evaluation and professional development systems with the potential for large salary increases, have improved the retention of high-performing teachers or increased attrition of low-performers (Cullen, Koedel, & Parsons, 2017; Dee & Wyckoff, 2015). Perhaps more surprisingly, some studies have found teachers responding similarly to relatively low-stakes evaluation reforms (Koedel, Li, Springer, & Tan, 2017; Sartain & Steinberg, 2016). However, much of the existing evidence comes from larger school systems and may not reflect the experiences of smaller districts or those with less capacity to administer more rigorous evaluations. One of the few national studies has found that the implementation of new state evaluation laws reduced the supply of newly licensed teachers (Kraft, Brunner, Dougherty, & Schwegman, 2018).

What is the issue?

A series of teacher evaluation reforms over the past decade have significantly altered the methods and consequences of teachers' evaluation reforms (McGuinn, 2015; Steinberg & Donaldson, 2016). The introduction of new performance measures was intended to address a perceived lack of rigor in existing systems, in which nearly all teachers were judged to be proficient (Kraft & Gilmour, 2017; Weisberg, Sexton, Mulhern, & Keeling, 2009). The failure of evaluations to meaningfully differentiate teachers' instructional effectiveness has been cited as an impediment to their use as a developmental tool and in support of accountability policies. Thus, in addition to incorporating new measures of teaching practice, states and districts have tied teacher evaluations more closely to accountability and professional development initiatives. Although there is little systematic evidence on the changing consequences of teacher evaluations, their stakes appear to vary considerably from place to place. State and district policies, for instance, are much more likely to include professional development requirements for low-performing teachers than to attach financial consequences to evaluations. Overall, about one fifth of the largest school districts in the United States provide bonus payments to teachers with high performance ratings (Steinberg & Donaldson, 2016). And high-profile reforms in New York City and Washington, D.C., have additionally linked tenure or dismissal decisions to the results of performance evaluations (Dee & Wyckoff, 2015; Loeb et al., 2015).

As states and districts reevaluate their policies with changes to federal regulations, it is worth revisiting the research base on the effects of evaluation policies on teacher and system outcomes. Given the broad reach of these reforms, the research literature covers an expansive set of questions related to teacher

evaluation, compensation, professional development, and turnover. In this brief, we consider a few topics where research has reached some conclusions about the implementation and effects of reforms, and others where the research is still tentative or incomplete.

What is known?

Quality of teacher evaluation metrics

The developmental and accountability goals of teacher evaluation systems both require accurate measures of teacher effectiveness. Thus, the reliability and validity of teacher evaluation metrics have been the focus of many studies, and evidence suggests that commonly used evaluation metrics are associated with student learning. Examining data on nearly 3,000 teachers, researchers affiliated with the Measures of Effective Teaching (MET) project found that common teacher effectiveness measures—classroom observations, value-added, and student surveys—predicted subsequent teacher performance, and that these predictions held up when teachers were randomized to classrooms (Kane, McCaffrey, Miller, & Staiger, 2013). Nonetheless, researchers have identified three primary concerns about using evaluation data for personnel decisions. Policymakers designing evaluation systems should be mindful of these limitations.

First, although teacher evaluation metrics are predictive of future performance on average, studies also show that individual teacher value-added is estimated imprecisely due to the relatively small number of students in each classroom. Persistent teacher effects account for about 50 percent of the variation for elementary teachers and 70 percent for middle school teachers (McCaffrey, Sass, Lockwood, & Mihaly, 2009). The variability of observational measures is similar (Mihaly et al., 2013). This variability means that teachers' performance ratings can vary substantially from year to year, potentially undermining trust in the evaluation system. Pooling data from multiple years or requiring several annual teacher observations can mitigate some of these problems (Goldhaber & Hansen, 2013; Kane & Staiger, 2012).

Second, teacher evaluation measures, particularly classroom observation measures, may reflect characteristics of students in the classroom rather than teacher quality. In one study of teachers in four school districts, teachers whose incoming students were higher achieving received higher classroom observation scores, on average, compared with teachers whose incoming students were lower achieving (Whitehurst, Chingos, & Lindquist, 2014). Analyzing random assignment data from the MET project, both Steinberg and Garrett (2016) and Campbell and Ronfeldt (2018) demonstrate that the academic performance of a teacher's incoming students affects teachers' classroom observation scores, such that teachers of low-performing students receive lower observation ratings. In addition, teachers in classrooms with high concentrations of Black, Hispanic, and male students receive lower observation ratings, and that these differences are unlikely due to actual differences in teacher quality (Campbell & Ronfeldt, 2018). Adjusting observational ratings for the effects of classroom composition ameliorates these sources of bias, although such adjustments reduce the transparency of these measures (Bacher-Hicks, Chin, Kane, & Staiger, 2017; Whitehurst et al., 2014).[1]

Finally, new research on the multidimensional nature of teaching suggests that existing measures fail to accurately capture other important teaching skills. For instance, although Chetty, Friedman, and Rockoff (2014) found that students assigned to a more effective teacher (based on value-added) are more likely to attend college and earn more, these results do not speak directly to the strength of the correlation between teachers' effects on test scores and teachers' effects on other outcomes. Subsequent analyses have suggested that these two types of teacher effects may not be strongly correlated. Chamberlain (2013) re-analyzed the Chetty, Friedman, and Rockoff data and found evidence that teachers' effects on test scores explain only a small proportion of their overall effect on students' educational attainment. Subsequent

research has found similar results in other locations, using other nontested outcomes, and for observational measures of teaching practice as well as teacher value-added (Blazar & Kraft, 2017; Gershenson, 2016; Jackson, 2018a; Kraft, 2017).[2]

Impact of evaluation systems on the effectiveness and composition of the teacher workforce

Several studies have documented that high-stakes evaluation reforms increase departure rates of low-performing teachers (Cullen et al., 2017; Dee & Wyckoff, 2015; Stecher et al., 2018). For example, in Houston a new evaluation policy increased the relatively likelihood of exit for low-performing teachers, and low-performing teachers in low-achieving schools were especially likely to leave. Similarly, under IMPACT (a multimeasure evaluation system in the District of Columbia Public Schools [DCPS]), receiving a rating that implied a strong dismissal threat increased voluntary attrition by 11 percentage points, or by more than 50 percent (Dee & Wyckoff, 2015). In New York City, which implemented a new tenure policy that allowed principals to extend the probationary period to allow teachers to improve and principals more time to gather information about performance, teachers whose probationary periods were extended were much more likely to leave their schools (Loeb, Miller, & Wyckoff, 2015). But there is also some evidence that reforms with lower stakes attached to evaluation results have affected teachers' career choices. Chicago's Excellence in Teaching Project pilot, which evaluated teachers multiple times using a classroom observation tool, increased exit for low-rated and nontenured teachers; teachers who exited were lower performing than the teachers who remained (Sartain & Steinberg, 2016). And Kraft et al. (2018) found that state evaluation reforms have reduced the supply of new teachers, even though these policies frequently attach fewer consequences to performance evaluations and frequently identify only a small number of teachers as low performing (Kraft & Gilmour, 2017; Steinberg & Donaldson, 2016).

Because increased teacher turnover may reduce student achievement, differential attrition is only likely to positively affect student outcomes if newly hired teachers are more effective than those they replace. Studies of evaluation reform efforts in DCPS, Chicago, New York City, and Wisconsin suggest that reforms that bundle evaluation with differential compensation can improve hiring outcomes (Adnot, Dee, Katz, & Wyckoff, 2017; Biasi, 2018; Loeb et al., 2015). In the New York City tenure reform and the DC IMPACT system, the replacement teachers were judged to be more effective than the teachers who exited (Adnot et al., 2017; Loeb et al., 2015). Similarly, a study of teacher pay reform in Wisconsin found that highly effective teachers working in districts where pay is based on seniority were more likely than low-quality teachers to move to a district that adopted flexible compensation, and high-quality teachers already teaching in flexible compensation districts were less likely to move to a seniority pay district compared with lower value-added teachers (Biasi, 2018). It is important to note, however, that improvements in hiring may not generalize to settings with weaker ties between compensation and previous classroom performance.

Despite these findings, there are still important unanswered questions about the overall effects of evaluation reforms on the teaching workforce. In particular, changes in attrition patterns or hiring alone may be insufficient to significantly improve overall teacher effectiveness. For instance, while the patterns documented by Adnot et al. (2017) and Dee and Wyckoff (2015) indicate that the DC IMPACT reforms likely led to an improvement in student learning, Cullen et al. (2016) suggest that changes in the Houston workforce following its evaluation reform were too small to have had a detectable impact on student achievement. Understanding the totality of labor market effects is crucial for evaluating the overall success of evaluation reforms, particularly if they affect the entry of new teachers into the profession.

In many locations, teacher evaluations support pay-for-performance programs offering bonuses to teachers based on their evaluations. These policies may encourage greater effort among existing teachers

without affecting teacher turnover or entry; however, findings from several randomized controlled trials suggest that such programs have not had the desired impact (Goodman & Turner, 2013; Marsh et al., 2011; Springer et al., 2010; Springer et al., 2012). For example, the Schoolwide Performance Bonus Program implemented in more than 200 New York City public schools had no effect on student achievement, attendance, or graduation (Fryer, 2013; Goodman & Turner, 2013; Marsh et al., 2011). Some evidence suggests that programs studied in these early experiments suffered from poorly designed incentives targeting groups rather than individual teachers. Goodman and Turner (2013) noted that while the Schoolwide Performance Bonus Program in New York had little effect on student achievement, effects on math achievement were more positive in schools where fewer teachers had responsibility for tested students (i.e. few "free riders"). Imberman and Lovenheim (2014) have found similar patterns under Houston's reforms.[3]

Developmental effects of evaluation reforms

Improving the practice of existing teachers is a central promise of teacher evaluation systems, and some studies have found that even low-stakes evaluation reforms affect student achievement. One of the first large-scale analyses of evaluation reforms came from the introduction of a low-stakes evaluation system for midcareer teachers in Cincinnati Public Schools. The Cincinnati policy was rolled out over several years, which allowed researchers to compare improvements for teachers evaluated during the first years of the policy to those evaluated later. Taylor and Tyler (2012) found that evaluations improved teachers' value-added by about 0.10 standard deviations, or significantly more than the difference between a novice and second-year teacher. A few studies have examined similar policy reforms in other districts; although the results have been more mixed, they generally suggest that low-stakes evaluations lead to improvements in teacher effectiveness (Garet et al., 2017; Steinberg & Sartain, 2016). Among the high stakes evaluation reforms, Dee and Wyckoff (2015) found that teachers who received low ratings that put them at threat of dismissal under DC IMPACT improved their performance by 0.27 of a teacher-level standard deviation. Notably, these effects all operate through changing the performance of existing teachers and suggest that evaluation reforms can improve teaching skill even without affecting turnover or career choice. Whether these findings reflect exceptional programs or effects of evaluation reforms that generalize to other settings remains an important area for future research.

What is not known?

How do evaluation policies support districts' development objectives?

The evidence on low-stakes reforms is promising and suggests that better feedback can drive improvements in teacher effectiveness. However, the studies cited in the prior sections assess bundles of policies rather than specific initiatives. We know less about which specific components of evaluation reforms explain these findings. In this section, we discuss several potential mechanisms by which low-stakes evaluation reforms might affect teacher effectiveness. Understanding the effects of these individual components of evaluation reforms—and the process by which they affect teacher outcomes—remains an important research objective.

Done well, teacher evaluation reforms increase the quantity and quality of information that teachers receive about their own performance. These reforms have increased the frequency and specificity of feedback to teachers about their performance (Donaldson & Woulfin, 2018; Steinberg & Donaldson, 2016; Sinnema & Robinson, 2007). But states and districts have also spent considerable effort communicating their expectations through descriptions of professional teaching standards and the development of observational rubrics. These initiatives may help establish a common language for describing classroom practices and facilitate conversations about effective teaching with peers and

administrators (Sartain et al., 2011). We do not know the importance of these effects, but if the shared understanding of effective teaching practices is an important component of evaluation reforms, states may want to consider other ways of communicating the content of professional standards to novice teachers. For instance, Massachusetts and Tennessee have made their state evaluation models part of the assessment process for teacher candidates' preservice clinical teaching experiences.

Evaluation reforms might also improve teacher practice by strengthening existing professional development initiatives. The most common link between evaluation and professional development is the use of performance ratings to target professional development toward low-performing teachers (Steinberg & Donaldson, 2016). Given the difficulty bringing effective coaching interventions to scale (Kraft et al., 2018), assigning professional development only to those with the lowest performance ratings might offer a cost-effective approach for districts to more intensively support struggling teachers. Further, if low-performing teachers benefit more from professional development activities, then using low-performance ratings as a screen might improve the overall effectiveness of these programs even while serving the same number of teachers. Thus far, we have only suggestive evidence on these points. In their analysis of Cincinnati's evaluation reform, Taylor and Tyler (2012) found that teacher performance improved only for teachers initially below the median performance level. This finding is consistent with the possibility that increasing professional development might especially benefit low-performing teachers, but other explanations are also possible. On the other hand, Stecher et al. (2018) reported no association between teachers' effectiveness and their perceptions of the professional development benefits of their districts' evaluation reforms. And Sartain et al. (2011) found that Chicago's evaluation reform only benefited teachers in low-poverty schools, which suggests that school climate may also moderate the effects of these programs. More evidence on how the effects of professional development vary by the measured effectiveness of teachers could inform the appropriateness of using evaluation results to screen teachers for additional training.

Some school systems also use evaluation results to tailor professional development to the individual needs of their teachers. As part of these initiatives, teachers' professional development depends in part on their performance on specific components of the evaluation framework. Examples of these initiatives include providing evaluation results to mentors to guide coaching activities and the development of individualized professional growth plans that address specific deficits. One outstanding question, however, is whether typical observational measures are reliable enough to support these policies. For instance, Garet et al. (2017) found that the year-to-year correlations in teachers' ratings on particular domains are significantly lower than the correlation in overall performance ratings. These findings call into question the utility of using performance ratings alone to identify domain-specific development activities for teachers. Nonetheless, there may be promise in more creative use of evaluation results. In a random assignment experiment in Tennessee, Papay, Taylor, Tyler, & Laski (2016) used state observational evaluation data to assign low-performing teachers higher performing mentors in the same school. They identified each teacher's lowest performance areas and matched them to mentors with strong performance ratings in those domains. Mentored teachers in schools who received these matching recommendations improved more on student achievement measures than teachers in schools using traditional approaches to recruit mentor teachers.

Does having both summative and formative purposes reduce the accuracy of evaluation scores?

Performance evaluations commonly serve both summative and formative purposes in new teacher evaluation systems. But attaching consequences to summative ratings may encourage administrators to adjust their ratings to avoid triggering certain sanctions. Even in low-stakes evaluation systems, providing low ratings to teachers can trigger additional documentation or training requirements for principals, which

might disincentivize those ratings (Kraft & Gilmour, 2017). Researchers have uncovered several apparently strategic responses by teachers and principals to the consequences of summative evaluations: School administrators tend to provide systematically higher performance ratings than neutral observers (Ho & Kane, 2013), give credit for professional contributions when assessing instructional effectiveness (Donaldson & Woulfin, 2018), and consider teachers' potential for future growth when assigning ratings (Kraft & Gilmour, 2017).

These adjustments likely reduce the quality of information about teaching practice embedded in performance evaluations. But we have less evidence about whether the combination of summative and formative evaluations undermines their benefit as a professional development tool. Even where principals artificially inflate formal ratings, they may still use the evaluation process to provide informal feedback that is more personalized and better differentiates the quality of instruction. Jacob and Lefgren (2008), for instance, found that principals' confidential ratings of teacher performance were more variable than the official ratings they reported to their district. It would, therefore, be informative to compare the professional development effects of teacher evaluation reforms, as in the studies described at the beginning of this section, in locations where the consequences vary.

How well do new measures capture other important components of teaching?

New evaluation systems incorporate many different instruments that reflect multiple dimensions of teaching skill. These include observational ratings of teacher practice, contributions to student achievement, and surveys of students. As we described above, there is a significant research literature that shows that these measures are correlated with one another and with other student academic outcomes. But recent research also indicates that teachers affect a variety of nontested student outcomes, such as attendance, discipline, educational attainment, and earnings. Most of the available evidence on the relationship between teacher evaluation metrics and nontest student outcomes comes from the value-added literature. But other indicators, such as student surveys, may be better positioned to capture teacher effects on students' noncognitive development.

Alternative measures of teacher effectiveness show promise as methods for expanding the set of teaching skills considered by researchers, but we currently know much less about how well these methods perform in practice as instruments for school or teacher evaluation. Thus far, policymakers have only begun to incorporate such measures into accountability systems, and few studies have assessed whether they accurately measure teacher effects or the extent to which they are influenced by classroom assignments (Duckworth & Yaeger, 2015; Loeb et al., 2018). In addition, when such measures rely on student outcomes directly, state administrative data systems may not measure the underlying outcomes as accurately as they do student achievement or teacher observational ratings (Murnane, 2013; National Forum on Education Statistics, 2018). Better understanding how—or whether—to incorporate other measures of student success into evaluation systems remains an important area of research.

Policy levers and policy-making challenges

Several policy levers can be used to increase the likelihood that teacher evaluation reforms will enhance the quality of the teaching workforce. First, policymakers should be mindful that reforms increasing the chances of teachers' losing their jobs may affect supply of new teachers or have unintended consequences on teacher turnover. Evaluation systems should create authentic opportunities for teachers to advance to offset these risks. Similarly, shifts away from using seniority for tenure may have unintended consequences if not paired with sufficient rewards or pay increases (Kraft et al., 2018; Rothstein, 2015). Evidence from DCPS's IMPACT suggests that substantial individual pay increases may help attract and

retain highly effective teachers in the context of evaluation reforms (Dee & Wyckoff, 2015), though caution is warranted in generalizing these findings to other school systems.

Policymakers should also think carefully about aligning the design of their evaluation programs with their professional development objectives. If policymakers intend for evaluation to improve teaching practice through observation and feedback, then it is important to ensure that schools and districts have sufficient capacity to use evaluation to drive instructional improvement. Principals and other school administrators have significantly increased their time spent observing and assessing teachers under new evaluation systems (Neumerski et al., 2018). Overburdened administrators may resort to a compliance orientation or artificially inflate ratings if giving low ratings would be too time-consuming for the rater (Kraft & Gilmour, 2017). School staff need adequate training, time, and resources to both conduct classroom observations and to implement required follow-up steps for low performers. At the school district level, coordination and alignment between offices overseeing evaluation, professional development, and curriculum may be essential to ensure program coherence.

Finally, policymakers confront challenges in designing evaluation systems that have both formative and summative uses. When evaluations are consequential, principals appear to compress ratings or engage in coaching that may inhibit their usefulness as sources of information about teaching practice. Although these practices may be an inevitable consequence of local leaders' role in interpreting policy mandates, policymakers may want to ensure that evaluation systems include opportunities for candid feedback about teaching practice. Ensuring these opportunities may require committing time and resources specifically to this purpose given the increased demands on principals' time under new evaluation systems.

References

Adnot, M., Dee, T., Katz, V., & Wyckoff, J. (2017). Teacher turnover, teacher quality, and student achievement in DCPS. *Educational Evaluation and Policy Analysis*, *39*(1), 54–76.

Bacher-Hicks, A., Chin, M. J., Kane, T. J., & Staiger, D. O. (2017). *An evaluation of bias in three measures of teacher quality: Value-added, classroom observations, and student surveys* (Working Paper No. 23478). Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w23478

Biasi, B. (2018). *The labor market for teachers under different pay schemes* (SSRN No. 2942134). Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.com/abstract=2942134

Blazar, D., & Kraft, M. A. (2017). Teacher and teaching effects on students' attitudes and behaviors. *Educational Evaluation and Policy Analysis, 39*(1), 146-170.

Campbell, S. L., & Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than we bargained for? *American Educational Research Journal, 55*(6), 1233-1267.

Chamberlain, G. (2013). Predictive effects of teachers and schools on test scores, college attendance, and earnings. *Proceedings of the National Academy of Sciences*, *110*(43), 17176–17182.

Chiang, H., Speroni, C., Herrmann, M., Hallgren, K., Burkander, P., & Wellington, A. (2017). *Evaluation of the Teacher Incentive Fund: Final report on implementation and impacts of pay-for-performance across four years* (No. NCEE 2018-4004). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from https://files.eric.ed.gov/fulltext/ED578857.pdf

Cullen, J. B., Koedel, C., & Parsons, E. (2016). *The compositional effect of rigorous teacher evaluation on workforce quality* (Working Paper No. 22805). Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w22805

Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, *34*(2), 267–297.

Donaldson, M. L., & Woulfin, S. (2018). From tinkering to going "rogue": How principals use agency when enacting new teacher evaluation systems. *Educational Evaluation and Policy Analysis*. Forthcoming.

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, *44*(4), 237–251.

Fryer, R. G. (2013). Teacher incentives and student achievement: Evidence from New York City Public Schools. *Journal of Labor Economics*, *31*(2), 373–407.

Garet, M. S., Wayne, A. J., Brown, S., Rickles, J., Song, M., & Manzeske, D. (2017). *The impact of providing performance feedback to teachers and principals* (No. NCEE 2018-4001). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from https://ies.ed.gov/ncee/pubs/20184001/pdf/20184001.pdf

Gershenson, S. (2016). Linking teacher quality, student attendance, and student achievement. *Education Finance and Policy*, *11*(2), 125–149.

Goldhaber, D., & Hansen, M. (2013). Is it just a bad class? Assessing the long-term stability of estimated teacher performance. *Economica*, *80*(319), 589–612.

Goodman, S. F., & Turner, L. J. (2013). The design of teacher incentive pay and educational outcomes: Evidence from the New York City bonus program. *Journal of Labor Economics*, *31*(2), 409–420.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel*. Seattle, WA: Bill and Melinda Gates Foundation. Retrieved from http://k12education.gatesfoundation.org/resource/the-reliability-of-classroom-observations-by-school-personnel/

Imberman, S. A., & Lovenheim, M. F. (2014). Incentive strength and teacher productivity: Evidence from a group-based teacher incentive pay system. *The Review of Economics and Statistics*, *97*(2), 364–386.

Jackson, C. K. (2018a). What do test scores miss? The importance of teacher effects on non–test score outcomes. *Journal of Political Economy, 126*(5), 2072-2107.

Jackson, C. K. (2018b). The full measure of a teacher. *Education Next*, *2019*(Winter), 63–68.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, *26*(1), 101–135.

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.* Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://k12education.gatesfoundation.org/resource/have-we-identified-effective-teachers-validating-measures-of-effective-teaching-using-random-assignment/

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Bill and Melinda Gates Foundation. Retrieved from http://k12education.gatesfoundation.org/resource/gathering-feedback-on-teaching-combining-high-quality-observations-with-student-surveys-and-achievement-gains-3/

Koedel, C., Li, J., Springer, M. G., & Tan, L. (2017). The impact of performance ratings on job satisfaction for public school teachers. *American Educational Research Journal*, *54*(2), 241–278.

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

Kraft, M. A. (2017). Teacher effects on complex cognitive skills and social-emotional competencies. *Journal of Human Resources*. Advance online publication. https://doi.org/10.3368/jhr.54.1.0916.8265R3

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.

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, *88*(4), 547–588.

Kraft, M. A., Brunner, E. J., Dougherty, S. M., & Schwegman, D. (2018). *Teacher accountability reforms and the supply of new teachers*. Unpublished manuscript. Retrieved from https://scholar.harvard.edu/files/mkraft/files/kraft_et_al._2018_teacher_accountability_reforms.pdf

Loeb, S., Christian, M. S., Hough, H., Meyer, R. H., Rice, A., & West, M. R. (2018). *School effects on social-emotional learning. Findings from the first large-scale panel survey of students.* Stanford, CA: Policy Analysis for California Education. Retrieved from https://www.edpolicyinca.org/publications/sel-school-effects

Loeb, S., Miller, L. C., & Wyckoff, J. (2015). Performance screens for school improvement the case of teacher tenure reform in New York City. *Educational Researcher*, *44*(4), 199–212.

Marsh, J. A., Springer, M. G., McCaffrey, D. F., Yuan, K., Epstein, S., Koppich, J., … Peng, A. (Xiao). (2011). *A big apple for educators: New York City's experiment with schoolwide performance bonuses (final evaluation report)*. Retrieved from https://www.rand.org/pubs/monographs/MG1114.html

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy*, *4*(4), 572–606.

McGuinn, P. (2015). *State education agencies and the implementation of new teacher evaluation systems*. Consortium for Policy Research in Education. Retrieved from http://www.cpre.org/sites/default/files/wp_mcguinn_2015.pdf

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). *A composite estimator of effective teaching*. Seattle, WA: Bill and Melinda Gates Foundation. Retrieved from http://k12education.gatesfoundation.org/resource/a-composite-estimator-of-effective-teaching/

Murnane, R. J. (2013). U.S. high school graduation rates: Patterns and explanations. *Journal of Economic Literature*, *51*(2), 370–422.

National Forum on Education Statistics. (2018). *Forum guide to collecting and using attendance data* (No. NFES 2017-007). Washington, DC: U.S. Department of Education, National Center for Education Statistics. Retrieved from https://nces.ed.gov/pubs2017/NFES2017007.pdf

Neumerski, C. M., Grissom, J. A., Goldring, E., Drake, T. A., Rubin, M., Cannata, M., & Schuermann, P. (2018). Restructuring instructional leadership: How multiple-measure teacher evaluation systems are redefining the role of the school principal. *The Elementary School Journal, 119*(2).

Papay, J. P., Taylor, E. S., Tyler, J. H., & Laski, M. (2016). *Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data* (No. 21986). Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w21986

Pham, L. D., Nguyen, T. D., & Springer, M. G. (2017, March). Teacher merit pay and student test scores: A meta-analysis. Presented at the Association for Education Finance and Policy Annual Meeting, Washington, DC.

Rothstein, J. (2015). Teacher quality policy when supply matters. *American Economic Review*, *105*(1), 100–130.

Rowan, B., Schilling, S. G., Spain, A., Bhandari, P., Berger, D., & Graves, J. (2013). *Promoting high quality teacher evaluations in Michigan: Lessons from a pilot of educator effectiveness tools*. Institute for Social Research, University of Michigan.

Sartain, L., & Steinberg, M. P. (2016). Teachers' labor market responses to performance evaluation reform: Experimental evidence from Chicago Public Schools. *Journal of Human Resources*, *51*(3), 615–655.

Sartain, L., Stoelinga, S. R., Brown, E. R., Luppescu, S., Matsko, K. K., Miller, F. K., … Glazer, D. (2011). *Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation*. Consortium on Chicago School Research. Retrieved from https://consortium.uchicago.edu/publications/rethinking-teacher-evaluation-chicago-lessons-learned-classroom-observations-principal

Sinnema, C. E. L., & Robinson, V. M. J. (2007). The leadership of teaching and learning: Implications for teacher evaluation. *Leadership and Policy in Schools*, *6*(4), 319–343.

Springer, M. G., Ballou, D., Hamilton, L. S., Le, V.-N., Lockwood, J. R., McCaffrey, D. F., … Stecher, B. M. (2010). *Teacher pay for performance: Experimental evidence from the Project on Incentives in Teaching*. Retrieved from https://www.rand.org/pubs/reprints/RP1416.html

Springer, M. G., Pane, J. F., Le, V.-N., McCaffrey, D. F., Burns, S. F., Hamilton, L. S., & Stecher, B. (2012). Team pay for performance: Experimental evidence from the Round Rock Pilot Project on Team Incentives. *Educational Evaluation and Policy Analysis*, *34*(4), 367–390.

Stecher, B. M., Holtzman, D. J., Garet, M. S., Hamilton, L. S., Engberg, J., Steiner, E. D., … Chambers, J. (2018). *Improving teaching effectiveness: The Intensive Partnership for Effective Teaching through 2015-2016, final report*. Santa Monica, CA: RAND Corporation. Retrieved from https://www.rand.org/pubs/research_reports/RR2242.html

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., & Sartain, L. (2015). Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching Project. *Education Finance and Policy*, *10*(4), 535–572.

Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *The American Economic Review*, *102*(7), 3628–3651.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect*. New York, NY: TNTP. Retrieved from https://tntp.org/publications/view/the-widget-effect-failure-to-act-on-differences-in-teacher-effectiveness

West, M. R., Buckley, K., Krachman, S. B., & Bookman, N. (2018). Development and implementation of student social-emotional surveys in the CORE Districts. *Journal of Applied Developmental Psychology*, *55*, 119–129.

Whitehurst, G. J., Chingos, M. M., & Lindquist, K. M. (2014, May). *Evaluating teachers with classroom observations: Lessons learned in four districts*. Washington, DC: Brown Center on Education Policy, Brookings Institute. Retrieved from http://www.brookings.edu/research/reports/2014/05/13-teacher-evaluation-whitehurst-chingos

[1] Value-added models typically control for a host of student characteristics to ameliorate these concerns (see, e.g., Koedel et al. [2015] for a discussion of commonly used student achievement metrics). This is not typically done for observational measures, although Whitehurst et al. (2014) suggest some methods for doing so.

[2] For a non-technical summary of this research, see Jackson (2018b).

[3] On the other hand, policymakers may see group incentives as a way of incentivizing collaboration among teachers (or offsetting the potential for individual performance pay to undermine collaboration). New York City's Schoolwide Performance Bonus may have had detrimental effects in schools with low levels of teacher cooperation and small positive effects on achievement in more cohesive schools (Goodman & Turner, 2013). Although it relies on both experimental and non-experimental analyses, at least one meta-analysis of pay-for-performance programs supports this hypothesis (Pham et al., 2017).