Evaluating and Rewarding Teachers

CASSANDRA HART

Abstract

Policymakers and researchers alike debate the optimal structure of teacher evaluation and compensation systems. This article reviews research in both fields, with a concentration on one increasingly policy-relevant topic in each domain. Within the evaluation domain, particular attention is given to value-added measures, which are increasingly being used to incorporate information about student test performance into teacher evaluations. While these measures allow evaluators to make quantitative estimates of teachers' contributions to student learning, critics argue that the measures suffer from a number of problems, including lack of stability, bias, and misattribution of teacher contributions. Within the realm of compensation, I devote particular attention to recent efforts to implement merit pay schemes, which aim to reward teachers, or teams of teachers, that are especially successful at boosting student achievement. Given that states and districts are increasingly requiring the use of value-added measures in evaluations and experimenting with merit pay plans, both areas are ripe for future research into the benefits and costs of these policies. Suggestions for future directions for research in both fields are offered.

INTRODUCTION

Educational quality has long been a primary concern for policymakers, and increasingly researchers and policymakers have looked to the role of teacher quality in promoting student achievement. While a broad consensus exists that teacher quality is perhaps the most important predictor of student achievement that is within the control of schools, there is significant controversy over how best to evaluate teachers to determine which teachers are providing the highest quality education, and whether to link teacher pay to evaluations.

Designing high-quality methods of teacher pay and evaluation are important for several reasons. The design of evaluation and pay systems may affect the level of effort that teachers put forth, as well as the incentives that teachers have to invest in improving their own teaching skills. Moreover, pay structure and evaluation methods are factors that potential teachers are likely to

Emerging Trends in the Social and Behavioral Sciences. Edited by Robert Scott and Stephen Kosslyn.
 ISBN 978-1-118-90077-2.

consider as they weigh whether to enter or remain in the profession. Evaluation and compensation systems therefore influence the composition of the teaching labor force, which is likely to have important implications for policy.

This essay reviews established knowledge about teacher pay and evaluation, including the current state of how teachers are paid and evaluated in the United States. It then turns to exciting, new threads of research in both areas, before concluding with broad suggestions about the likely future of the field.

FOUNDATIONAL RESEARCH

Compensation

For the vast majority of teachers, pay is determined by the "single salary schedule," a district-specific (or state-specific) formula that dictates how much teachers will earn based on factors such as the number of years they have been teaching and the highest degree held (Odden & Kelley, 2001). As of 2003–2004, 96% of districts adhered to the single-salary schedule (Podgursky, 2009).

Critics of the single-salary schedule contend that the lack of connection between teachers' performance and their compensation disincentivizes educators from putting forth their maximal effort (Hanushek, 1981). Underlying the debate over how best to compensate teachers is a wealth of theoretical literature in economics and industrial organization. This literature suggests that as a general rule, tying employees' pay to their output should motivate them to work harder (Lazear, 2000). However, theory also suggests that designing performance pay structures is particularly difficult in complex professions and in professions where teamwork is important; both of these conditions hold for teaching (Holmstrom & Milgrom, 1991; Murnane & Cohen, 1986). For instance, society may expect teachers to turn out students who are not only competent in core subjects but also socialized to be conscientious, goal-oriented, curious, and civically engaged. At the same time, students may benefit from both individual teacher efforts in the class, and efforts among teams of teachers (e.g., to coordinate instruction between teachers of different subjects, or to share information among teachers within the same subject about effective instructional strategies for particular concepts); payment schemes based on individual performance may undermine such teamwork (Murnane & Cohen). Because it is difficult to parse out the responsibility for student success and because it is complicated to define what student success means, paying teachers based on their performance in this domain is especially difficult. Both of these areas have been contested in the literature on optimal methods of teacher evaluation.

Evaluation

Historically, the vast majority of school systems have relied on principal evaluations of teacher performance. However, principal incentives to judge teachers stringently are minimal, especially once teachers have already received tenure. This has resulted in unrealistically lax standards; for instance, less than 1% of teachers in the Chicago Public School system received unsatisfactory ratings from the 2003–2004 to the 2007–2008 school year; over 90% were judged "Superior" or "Excellent" (Weisberg, Sexton, Mulhern, & Keeling, 2009). Moreover, while principals are successful at identifying the teachers who promote the highest and lowest student achievement gains, they are less able to distinguish between teachers in the middle of the distribution (Jacob & Lefgren, 2008).

These problems with traditional evaluation systems have prompted calls for more rigor in teacher evaluation. Two main methods have been proposed: using richly detailed observations of teachers' practice and using student test scores to measure teacher quality. The former method generally involves highly trained observers evaluating teachers using a specific rubric to quantify the quality of teacher practice. Often, such systems employ multiple evaluators, including professional observers independent of the school, to ensure greater objectivity than is provided by traditional principal evaluations. While such systems have historically been the dominant form of "objective" evaluation, in recent years reformers have increasingly looked to use "value-added" measures as a less resource-intensive, and more easily quantifiable, way to evaluate teacher performance.

Value-added measures use student test scores and attempt to identify the unique contribution that an individual teacher makes to boosting student achievement in a given year. In effect, value-added measures compare the test score gains a teacher's students actually make over their own year-prior performance, to the gains that would have been expected for those students if they had been taught by a statistically "average" teacher. These methods have become more appealing as states have developed testing regimes that test students every year in accordance with either state or federal accountability policies. Early studies in value-added measures suggest that teacher quality varies widely among individual teachers, and that teacher quality is important for improving student achievement (Rivkin, Hanushek, & Kain, 2005; Wright, Horn, & Sanders, 1997). A one standard deviation increase in the quality of the teacher as measured by value-added scores is estimated to produce a level of benefit similar to a 10-pupil reduction in class size (Rivkin, Hanushek, & Kain). Notably, the relationship between value-added measures and certain formal qualifications used to determine salaries is weak. For instance, completion of an advanced degree is not

strongly associated with student achievement (Rivkin, Hanushek & Kain; Harris & Sass, 2011; but see Clotfelter, Ladd, & Vigdor, 2007). And while novice teachers produce smaller achievement gains than teachers with more experience, researchers find little additional benefit to experience past the first few years (Nye, Konstantopoulos, & Hedges, 2004; Rivkin, Hanushek & Kain; but see Clotfelter, Ladd, & Vigdor).

Although value-added measures have the advantage of providing "objective" feedback on teacher quality, they have been criticized along several dimensions as well. The question of how best to construct value-added measures has been fraught. For instance, researchers debate whether to control for student characteristics. Failing to control for student characteristics might mean that teachers are effectively punished for teaching populations that face greater challenges in school (e.g., English learners or low-income students). On the other hand, controlling for student background is politically unpalatable, suggesting that all students are not equally able to learn (Ballou, Sanders, & Wright, 2004). Another alternative is to include student fixed effects, which control for time-invariant unobserved student characteristics, but which place heavy computational demands on the data and increase the sensitivity of the measures to model specifications (Harris, Sass, & Semykina, 2010). Researchers and policymakers similarly question which school characteristics to control for: Should factors within districts' control, such as class size, factor into value-added measures? Should the school itself be controlled for, so that teachers are effectively only compared to colleagues within the same school, or does that suggest an acceptance of inequality in teacher effectiveness between schools? Models may come to different conclusions about teacher effectiveness depending on the factors that are controlled (Ballou, Sanders, & Wright).

Furthermore, value-added measures have been criticized for a lack of stability. Researchers may reach different conclusions about a teacher's effectiveness if they look at two different years of data, although this temporal instability can be addressed by using multiple years of data (McCaffrey, Sass, Lockwood, & Mihaly, 2009). Teacher value-added estimates also vary based on the test used (Papay, 2010), and even within the same test, teacher effectiveness looks different depending on the subject subdomain examined (Lockwood *et al*., 2007).

Value-added measures have also come under attack for bias. For instance, a key assumption of the measures is that students are randomly assigned to teachers. However, this assumption is often violated in practice (Rothstein, 2010), although the degree of this bias is reduced by using information from several years of measures (Koedel & Betts, 2011).

In addition, while value-added measures ideally isolate the contribution of a single teacher to a student's test scores, they likely reflect the efforts of

several teachers. For instance, if achievement tests are given in March, the March-to-March change in a student's performance will reflect both the contributions of this year's teacher (from September to March) and last year's teacher (from March to June). Moreover, it will include any summer learning that the student achieved through summer school or informal learning experiences at home, camp, or elsewhere (Papay, 2010). Likewise, in middle and high schools, where students are assigned to different teachers for different subjects, there may be spillover effects through which, say, one's math teacher affects reading scores; the evidence on this phenomenon is mixed (Koedel, 2009). All of these problems have led to questions over the validity of value-added measures.

CUTTING-EDGE RESEARCH

Evaluation

A wealth of recent studies has extended researchers' understandings of the strengths and weaknesses of the use of value-added measures for teacher evaluation. A particularly important new development in this field has been the validation of teacher value-added measures with data from randomized control trials. While value-added estimates use statistical techniques to try to adjust for factors such as composition of the class to the greatest extent possible, researchers remained concerned that student–teacher matching based on unobserved characteristics (such as motivation or family involvement) might drive results. Kane and Staiger (2008) address this gap in the literature by working with a large urban district to randomly assign students to teachers who had different levels of value-added scores calculated by traditional statistical methods in the previous year. If value-added scores are unbiased, the students randomly assigned to the teachers with the higher historical value-added scores would be expected to outperform their peers assigned to a historically lower value-added teacher. In fact, this was what the researchers observed, suggesting that value-added measures provide an unbiased measure of teacher quality when prior student achievement is controlled (Kane & Staiger).

This work has been extended by the Measures of Effective Teaching project (Cantrell & Kane, 2013), which uses value-added measures in combination with teacher evaluations and student surveys to provide a multidimensional view of teacher quality. Determining the predictive value of such multidimensional measures is important because they are more likely to be actually implemented by districts than are value-added measures alone, given that value-added measures are politically contentious and cannot reasonably capture all aspects of a teacher's performance. The MET project finds

that these multidimensional measures are predictive of student achievement under random assignment.

While these studies show that value-added measures are successful at predicting improvements in student achievement, other researchers have found that value-added measures are also useful for predicting student outcomes in other domains. Assignment to high-value-added high school math teachers is associated with a greater likelihood of graduation (Koedel, 2008). And recent studies find long-term gains for students assigned to high-value-added teachers in the primary grades, as evidenced by higher earnings in adulthood and improved college outcomes (Chetty *et al.*, 2011). These studies suggest that teachers who produce better achievement also have positive effects for a range of other outcomes that policymakers want to promote.

At the same time, value-added measures that rely on test scores alone may fail to identify some teachers who are particularly good at boosting students' noncognitive skills. For instance, one new study identifies a noncognitive factor that, controlling for student achievement, is associated with student outcomes such as grade progression, suspension rates, and absences (Jackson, 2012). Teachers have important effects on this noncognitive dimension, and it is imperfectly captured by value-added measures of achievement. This suggests that traditional value-added measures based on achievement alone may not identify teachers who are particularly good at fostering noncognitive skills that are also linked to important adult outcomes such as earnings or arrests.

Other recent work adds another interesting caveat to the use of value-added measures: Teachers may not be equally effective with all students or in all settings. Recent work suggests that between 10–40% of what is estimated as teacher quality can be explained by match quality between the teacher and the school (Jackson, 2013). Moreover, interactions between the teacher and individual students matter somewhat as well, accounting for about 3–4% of the variance in teacher effects in different classes (Lockwood & McCaffrey, 2009). This line of work is important because it suggests that teacher quality is not fully portable across settings and with all students.

COMPENSATION

A number of important studies of teacher compensation have complemented these studies on evaluation. In particular, there have been several recent experiments that have implemented pay-for-performance schemes of the type that some theorists contend should motivate teachers to increase their effort. The findings of these evaluations within the United States have not been encouraging. The majority of performance pay programs, in

areas as diverse as Tennessee (Springer *et al*., 2010), Chicago (Glazerman, McKie, & Carey, 2009), and New York (Springer & Winters, 2009; Fryer, 2011) have shown either null or very small and inconsistent effects on student performance. Bonus sizes for these interventions ranged from about $2,000 to $15,000 per year.

One interesting exception employs the power of loss aversion to increase the salience of the teacher incentives. Psychologists have long known that people are more motivated to avoid losses than they are to achieve gains of an equivalent amount (Kahneman & Tversky, 1984). Harnessing this insight, researchers randomly assigned teachers to one of two bonus conditions (Fryer, Levitt, List, & Sadoff, 2012). The first group (the "Gain" group) was eligible for an $8000 bonus to be paid at the end of the year if their students met performance target. The second group (the "Loss" group) received a $4000 bonus payment upfront, which was revoked if their students failed to meet performance goals. If their students met the performance targets, teachers in the "Loss" group would keep the initial payment and receive an additional $4000 year-end bonus. While both groups stood to gain identical amounts for meeting performance targets, student achievement improved significantly more for teachers in the Loss group. This was true whether bonuses were assigned on the basis of individual or team performance. This intervention suggests that changes in the framing of bonus policies can affect their efficacy and points to an interesting new direction for future research.

Although these experiments have not been large enough to affect teacher labor supply on a large scale, some theoretical work has begun to tackle the question of how teacher labor supply may be affected by linking pay and retention to teacher performance. Examining the possible effects of determining firing decisions based on performance rather than seniority through simulations, Boyd and colleagues (2011) find evidence that tying retention to performance would improve the overall quality of the teaching force. However, other simulation studies caution that the efficacy of such policies may be mitigated when the measures used to judge teachers are subject to manipulation (Rothstein, 2012).

KEY ISSUES FOR FUTURE RESEARCH

Research on teacher evaluation and compensation systems should blossom in the coming years as states and localities increasingly experiment with policies intended to better motivate teachers. Experimental research will be one important part of the research puzzle: Experiments are necessary to help policymakers determine the compensation and retention frameworks that best promote improvement on student test scores, and on other important dimensions that policymakers care about, such as graduation and college-going

behavior. These experiments would likely involve randomizing school participation in different forms of incentive schemes, as past experiments in the pay-for-performance literature have done. As such, these would take place on a relatively small scale, to be scaled up when states have found incentive frameworks that seem to optimize their defined goals.

To complement experimental studies, quasi-experimental research will be necessary to evaluate, at a larger scale, the efforts of states and districts to incorporate value-added measures into evaluation and compensation decisions. Several states, such as Florida, Tennessee, and Rhode Island, have started to enact such measures already, although these new policies face litigation in some states, including Florida (National Council on Teacher Quality, 2011). Given that these policies are enacted to incentivize teachers to improve students' academic performance, a crucial question will be how changes to evaluation and compensation systems affect test scores. However, a more complete reckoning will also include a complement of non-test-score measures. Researchers should examine whether teacher effects on outcomes such as attendance, graduation, and disciplinary actions change as teacher-level accountability for test scores is added.

In addition to changing how current teachers perform in their jobs, introducing accountability for test scores at the teacher level is likely to affect the composition of the teacher labor force. This introduces a suite of questions for researchers to answer. Do systems that compensate for performance increase the quality of incoming recruits to the teaching labor force, or does the lack of predictability in compensation repel qualified potential teachers? How is teacher turnover affected by these policies? Researchers should seek to establish how the overall quality of the teaching force is affected by policy changes that tie compensation and evaluation to test scores, and the points at the pipeline at which any changes in overall teacher quality occur.

Studying the effects on distribution of teachers among different types of students will also be critical. In theory, adjusting for students' prior-year test scores should ensure that teachers who are assigned to lower achieving students are not penalized for this assignment. However, because year-to-year changes incorporate not only school-year learning rates, but the rate of learning incurred over the summer months, teachers who are held accountable for student learning may be incentivized to avoid teaching students with greater rates of learning loss over the summer. On average, summer learning loss is more acute for students of low socioeconomic status (Downey, von Hippel, & Broh, 2004); if teachers jockey to avoid teaching low-income students in order to maximize their compensation or minimize their likelihood of dismissal, evaluation policies could impose an unintended cost on these disadvantaged students.

On a similar note, a promising direction for future research is to examine how class composition affects various aspects of teacher value-added measures. While the inclusion of student-level covariates may protect value-added measures from *bias* associated with class composition, it is an open question whether the *stability* of the measures is affected. Evidence from the accountability literature suggests, for instance, that achievement scores of English language learners are less stable than those of native English speakers (Abedi, 2004); teaching classes with large concentrations of students with predictably less stable scores should make teacher value-added scores less stable as well. Teachers with less stable measures of value-added effects will be more likely to be misclassified as either high- or low-performing; this ramification of unstable measures is a well-known problem in the school accountability literature (Kane & Staiger, 2002).

Given these concerns, a particularly important area of study will be to examine the effects of changes in teacher evaluation and compensation policies on classroom experiences of students and teachers, and on overall school climate. A healthy body of literature has documented that teachers spend more time on tested subjects, and on tested concepts within a given subject, when school-level accountability is introduced (McMurrer, 2007; Srikantaiah, Zhang, & Swayhoover, 2008). Researchers should examine whether the tendency to increase emphasis on tested subjects and concepts is heightened further when teacher evaluation and compensation decisions are tied to those subjects.

Quantitative studies that address these questions must be complemented by high-quality qualitative work. Qualitative work on school-level accountability has revealed a number of important insights, including techniques that school administrators under new accountability systems used to "game the system" with potentially adverse educational effects (e.g., Booher-Jennings, 2005). Examples include encouraging teachers to focus their attention on students on the bubble of passing proficiency thresholds on standardized tests, while effectively ignoring students that teachers consider almost certain to fail (or certain to pass) (Booher-Jennings). Similar work should examine the effects of changes in teacher compensation and evaluation systems. Extensive interviews should be conducted with current teachers regarding the effects of the introduction of value-added measures on their effort level, morale, interaction with colleagues, and career plans. Similar interviews should also be carried out with teachers who have left the profession to pursue other career options. Qualitative work should extend past current and former teachers to include potential future teachers as well. College students in states with different evaluation and compensation policies should be interviewed regarding their awareness of these policies and the effect changes in these policies would have on their propensity to

enter the teaching field. It is particularly important to interview students in states that have not yet adopted value-added measures as a component of evaluation and compensation to determine how potential teachers' responses change across cohorts as state policy changes.

Interviews of current, former, and would-be teachers should be complemented by rich classroom observations of teachers under different compensation and evaluation systems. Such observations can determine whether classroom practice differs, for instance, among novice teachers under systems that offer tenure versus a series of one-year contracts; again, it would be particularly useful to conduct such studies over time in states or districts that are likely to implement changes to see if there are changes within the same district under different systems.

The breadth of questions raised by changes to evaluation and compensation system demands attention from researchers from multiple disciplines. Statisticians and economists can both contribute to the work surrounding the best structure for value-added measures. At the same time, as the experiments by Fryer and colleagues show (2012), insights drawn from behavioral economics and psychology may be useful in determining how best to structure compensation and evaluation programs. Quantitative analysts who use both quasi-experimental and experimental methods will be able to bring different perspectives to bear in evaluating the impact of policies that put value-added and other evaluation systems to use. And psychologists and sociologists should be encouraged to study likely effects on individual teachers' motivation, school organization, and school cohesion. Crucially, all of these researchers should engage with teachers, principals, and superintendents to ensure that research is aligned with the concerns of those who will be in the classrooms.

Another challenge associated with the compensation and retention questions in particular is that many of them will require a relatively longer timeframe to study. While tying compensation to performance may have effects on test scores in the short-term, effects on factors such as the composition of the workforce may be take a longer time to become evident, and may change as time passes. For instance, changes in compensation may not change the plans of students who are nearing their college graduation, but may change the attractiveness of teachers to the current crop of high school students. This field will therefore demand the attention of researchers for years to come.

ACKNOWLEDGMENTS

Thanks to Heather Rose for comments on a draft of this essay. Any errors are my own.

REFERENCES

Abedi, J. (2004). The no child left behind act and english language learners: Assessment and accountability issues. *Educational Researcher*, *33*(1), 4–14. doi:10.3102/0013189X033001004

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessments of teachers. *Journal of Educational and Behavioral Statistics*, *29*(1), 37–65. doi:10.3102/10769986029001037

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas Accountability System. *American Educational Research Journal*, *42*(2), 231–268. doi:10.3102/00028312042002231

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. doi:10.1162/EDFP_a_00041

Cantrell, S., & Kane, T. J. (2013). Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET Project's three-year study. Measures of Effective Teaching Policy and Practitioner Brief. Seattle, WA: Bill & Melinda Gates Foundation. Accessed 2/13/2012. Retrieved from http://metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *Quarterly Journal of Economics*, *126*(4), 1593–1660. doi:10.1093/qje/qjr041

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, *26*, 673–682. doi:10.1016/j.econedurev.2007.10.002

Downey, D. B., von Hippel, P. T., & Broh, B. A. (2004). Are schools the great equalizer? Cognitive inequality during the summer months and the school year. *American Sociological Review*, *69*(5), 613–635. doi:10.1177/000312240406900501

Fryer, R. (2011). Teacher incentives and student achievement: Evidence from New York City Public Schools. NBER working paper 16850. Cambridge, MA: National Bureau of Economic Research. Accessed 2/13/2013. Retrieved from http://www.nber.org/papers/w16850.

Fryer, R. G., Levitt, S. D., List, J., & Sadoff, S. (2012). Enhancing the efficacy of teacher incentives through loss aversion: A field experiment. NBER working paper 18235. Cambridge, MA: National Bureau of Economic Research. Accessed 2/12/2013. Retrieved from http://www.nber.org/papers/w18237.

Glazerman, S., McKie, A., & Carey, N. (2009). Evaluation of the Teacher Advancement Program (TAP) in the Chicago Public Schools: Study design report. Document No. PR08-14. Washington, DC: Mathematica Policy Research. Accessed 2/27/2013. Retrieved from http://www.mathematica-mpr.com/publications/pdfs/education/TAP_rpt.pdf

Hanushek, E. A. (1981). Throwing money at schools. *Journal of Policy Analysis and Management*, *1*(1), 19–41. doi:10.2307/3324107

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality, and student achievement. *Journal of Public Economics*, *95*(7–8), 798–812. doi:10.1016/j.jpubeco.2010.11.009

Harris, D., Sass, T., & Semykina, A. (2010). Value-added models and the measurement of teacher productivity. National Center for the Analysis of Longitudinal Data in Education Research Working Paper 54. Washington, DC: Urban Institute. Accessed 2/12/2013. Retrieved from http://www.urban.org/UploadedPDF/1001508-Measurement-of-Teacher-Productivity.pdf.

Holmstrom, B., & Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization*, *7*, 24–52.

Jackson, C. K. (2012). Non-cognitive ability, test scores, and teacher quality: Evidence from 9th grade teachers in North Carolina. NBER working paper 18624. Cambridge, MA: National Bureau of Economic Research. Accessed 2/12/2013. Retrieved from http://www.nber.org/papers/w18624.

Jackson, C. K. (2013). Match quality, worker productivity, and worker mobility: Direct evidence from teachers. *Review of Economics and Statistics*, *95*(4), 1096–1116.

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. doi:10.1086/522974

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, *39*(4), 341–350. doi:10.1037/0003-066X.39.4.341

Kane, T. J., & Staiger, D. O. (2002). The promises and pitfalls of using imprecise school accountability measures. *The Journal of Economic Perspectives*, *16*(4), 91–114. doi:10.1257/089533002320950993

Kane, T. J., & Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. NBER working paper 14607. Cambridge, MA: National Bureau of Economic Research. Accessed 2/12/2013. Retrieved from http://www.dartmouth.edu/~dstaiger/Papers/WP/2008/KaneStaiger%20NBER%20wp14607%202008.pdf.

Koedel, C. (2009). An empirical analysis of teacher spillover effects in secondary school. *Economics of Education Review*, *28*, 682–692. doi:10.1016/j.econedurev.2009.02.003

Koedel, C. (2008). Teacher quality and dropout outcomes in a large, urban school district. *Journal of Urban Economics*, *65*, 560–572. doi:10.1016/j.jue.2008.06.004

Koedel, C., & Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Policy*, *6*(1), 18–42. doi:10.1162/EDFP_a_00027

Lazear, E. (2000). Performance pay and productivity. *American Economic Review*, *90*(5), 1346–1361. doi:10.1257/aer.90.5.1346

Lockwood, J. R., & McCaffrey, D. F. (2009). Exploring student-teacher interactions in longitudinal achievement data. *Education Finance and Policy*, *4*(4), 439–467. doi:10.1162/edfp.2009.4.4.439

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, *44*(1), 47–67. doi:10.1111/j.1745-3984.2007.00026.x

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. doi:10.1162/edfp.2009.4.4.572

McMurrer, J. (2007). *Choices, changes, and challenges: Curriculum and instruction in the NCLB era*. Washington, DC: Center on Education Policy.

Murnane, R. J., & Cohen, D. K. (1986). Merit pay and the evaluation problem: Why most merit pay plans fail and a few survive. *Harvard Educational Review*, *56*(1), 1–18.

National Council on Teacher Quality (2011). *State of the states: Trends and early lessons on teacher evaluation and effectiveness policies*. Washington, DC: Author. Accessed 2/13/2013. Retrieved from http://www.nctq.org/p/publications/docs/nctq_stateOfTheStates.pdf.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*(3), 237–257. doi:10.3102/01623737026003237

Odden, A., & Kelley, C. (2001). *Paying teachers for what they know and do: New and smarter compensation strategies to improve schools*. Thousand Oaks, CA: Corwin Press.

Papay, J. (2010). Different tests, different answer: The stability of value-added estimates across outcome measures. *American Education Research Journal*, *48*(1), 163–193. doi:10.3102/0002831210362589

Podgursky, M. (2009). Market based pay reforms for teachers. In M. G. Springer (Ed.), *Performance incentives: Their growing impact on American K-12 education* (pp. 67–86). Brookings Institution Press: Washington, DC.

Rivkin, S., Hanushek, E., & Kain, J. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–58. doi:10.1111/j.1468-0262.2005.00584.x

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, *125*(1), 175–214. doi:10.1162/qjec.2010.125.1.175

Rothstein, J. (2012). Teacher quality policy when supply matters. NBER working paper 18419. Cambridge, MA: National Bureau of Economic Research. Accessed 2/12/2013. Retrieved from http://www.nber.org/papers/w18419.

Springer, M. G., Ballou, D., Hamilton, L., Le, V., Lockwood, J. R., McCaffrey, D. F., … Stecher, B. M. (2010). Teacher pay for performance: Experimental evidence from the Project on Incentives in Teaching. Nashville, TN: National Center for Performance Incentives. Accessed 2/13/2013. Retrieved from https://my.vanderbilt.edu/performanceincentives/files/2012/09/POINT_REPORT_9.21.102.pdf.

Springer, M. G., & Winters, M. (2009). New York City's School-wide Bonus Pay Program: Early evidence from a randomized control trial. National Center for Performance Incentives working paper 2009-02. Nashville, TN: National Center for Performance Incentives. Accessed 2/13/2013. Retrieved from

https://my.vanderbilt.edu/performanceincentives/ncpi-publications/program-evaluations-and-experiments/new-york-city-evaluation/new-york-citys-school-wide-bonus-pay-program-early-evidence-from-a-randomized-trial/.

Srikantaiah, D., Zhang, Y., & Swayhoover, L. (2008). *Lessons from the classroom level: Federal and state accountability in Rhode Island*. Washington, DC: Center on Education Policy.

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

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, *11*, 57–67. doi:10.1023/A:1007999204543

FURTHER READING

Corcoran, S. P. (2010). *Can teachers be evaluated by their students' test scores? Should they be? The use of value-added measures of teacher effectiveness in policy and practice*. Providence, RI: Annenberg Institute for School Reform. Retrieved from http://steinhardt.nyu.edu/scmsAdmin/uploads/006/265/valueAddedReport.pdf.

Hanushek, E. A., & Rivkin, S. G. (2012). The distribution of teacher quality and implications for policy. *Annual Review of Economics*, *4*, 131–157.

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

Odden, A., & Kelley, C. (2001). *Paying teachers for what they know and do: New and smarter compensation strategies to improve schools*. Thousand Oaks, CA: Corwin Press.

Podgursky, M., & Springer, M. (2011). Teacher compensation systems in the United States K-12 public school system. *National Tax Journal*, *64*(1), 165–192.

CASSANDRA HART SHORT BIOGRAPHY

Cassandra Hart is an assistant professor at the University of California, Davis School of Education. She conducts work examining the effects on student outcomes of state and national education policies. Her most recent work has focused on school choice. Hart earned her PhD from the Department of Human Development and Social Policy at Northwestern University in 2011.

RELATED ESSAYS

Economics of Early Education *(Economics)*, W. Steven Barnett
Shadow Education *(Sociology)*, Soo-yong Byun and David P. Baker
Misinformation and How to Correct It *(Psychology)*, John Cook *et al*.

Four Psychological Perspectives on Creativity *(Psychology)*, Rodica Ioana Damian and Dean Keith Simonton
The Organization of Schools and Classrooms *(Sociology)*, David Diehl and Daniel A. McFarland
Expertise *(Sociology)*, Gil Eyal
Controlling the Influence of Stereotypes on One's Thoughts *(Psychology)*, Patrick S. Forscher and Patricia G. Devine
Evolutionary Approaches to Understanding Children's Academic Achievement *(Psychology)*, David C. Geary and Daniel B. Berch
The Evidence-Based Practice Movement *(Sociology)*, Edward W. Gondolf
Educational Testing: Measuring and Remedying Achievement Gaps *(Educ)*, Jaekyung Lee
Retrieval-Based Learning: Research at the Interface between Cognitive Science and Education *(Psychology)*, Ludmila D. Nunes and Jeffrey D. Karpicke
The Impact of Learning Technologies on Higher Education *(Psychology)*, Chrisopher S. Pentoney *et al.*
Curriculum as a Site of Political and Cultural Conflict *(Sociology)*, Fabio Rojas
Education in an Open Informational World *(Educ)*, Marlene Scardamalia and Carl Bereiter
Leadership *(Anthropology)*, Adrienne Tecza and Dominic Johnson