INTRODUCTION

At least since the release of the Coleman Report in 1966, economists have been interested in

their type of teaching certification.[9] However, early studies document benefits to additional years of experience for early career teachers (e.g., Rockoff 2004, Rivkin et al. 2005), and recent studies indicate that teacher effectiveness meaningfully improves well into a teacher's career (Papay & Kraft 2013, Wiswall 2013).[10] We return to this evidence below in our discussion of policies to improve teacher effectiveness, such as mentoring and feedback.

Two recent studies depart from the use of administrative data in search of a stronger link between teacher effects and teacher characteristics. Rockoff et al. (2011) conduct an online survey of new elementary and middle school math teachers in New York City to collect data on characteristics linked to performance in similar occupations, such as general intelligence, personality traits, and beliefs regarding self-efficacy. They find that no single characteristic is a strong predictor of student achievement growth. However, they can predict roughly 10% of the variation in new teacher performance using indices that combine these characteristics.

Dobbie (2011) takes a similar approach, using data from Teach for America (TFA), in which new teaching candidates are rated by TFA admissions staff on eight criteria used to make program selection decisions. Scores on each criterion (academic achievement, leadership experience, perseverance, critical thinking, organizational ability, motivational ability, respect for others, and commitment to the TFA mission) are based on information collected from an online application, a phone interview, and an in-person interview. In line with Rockoff et al.'s (2011) previous findings, Dobbie finds that few of these TFA criteria are individually significant predictors of student achievement growth, but they strongly predict new TFA teacher performance when averaged into an index. His results suggest that more than half of the variance in the value-added of TFA teachers could be predicted based on their admission scores.[11]

Thus, it appears that some headway may be made in identifying the type of individuals who are likely to succeed in teaching through more intensive and purposeful data collection during the

In addition, Chetty et al. (2013a) implement a quasi-experimental test for bias, using changes in the mean teacher effect at the cohort level and changes in achievement across cohorts. Intuitively, if a fourth-grade teacher with low (estimated) value-added leaves a school and is replaced by a teacher with high (estimated) value-added, we would expect that average fourth-grade achievement would rise. Moreover, if the teacher effect estimates are unbiased, cohort-level scores should rise by the difference in the two teachers' effects, multiplied by the share of students they taught.[12] In contrast, if the two teachers were actually equally effective (i.e., their estimated effects were driven purely by sorting), then there should be no cross-cohort change in scores.

Using data from a large urban school district, Chetty et al. (2013a) find no evidence for this bias: the quasi-experimental cross-cohort variation in teacher effects is an unbiased predictor of cross-cohort changes in achievement. **Figure 2** illustrates this finding. **Figure 2*a*** is not the quasi-experiment but is presented for purposes of comparison. It shows student achievement residuals in a teacher's current classroom plotted against the teacher's value-added estimate based on different students in other years. There is a very clean one-to-one relationship, as expected based on the construction of value-added. As discussed above, the persistence of value-added over time could be driven by bias as well as through causal impacts of individual teachers. The quasi-experimental test is shown in **Figure 2*b***, which plots cross-cohort changes in student achievement against cross-cohort changes in teacher value-added. The one-to-one relationship holds up strikingly well; test score changes and value-added changes are tightly linked, even when looking only across adjacent cohorts of students within the same school and grade. In addition to a number of other checks provided by Chetty et al. (2013a), the relationship shown in **Figure 2*b*** strongly supports the notion that there is minimal predictive bias in their value-added measures.

Of course, there can be no guarantee that teacher effects estimated in other samples will be similarly unbiased. Because teachers are not randomly assigned, the properties of teacher effect estimates will depend on the quality of control variables that account for differences across students. However, it appears that the data and methods most commonly applied in this field are able to establish a causal link between teachers and student achievement.

Although the evidence on bias is supportive of teacher effects, stability may be even more important in determining accuracy in predicting teachers' future effects. Sizeable student- and class-level error components mean that a teacher effect based on just one or two classrooms can be a noisy indicator of a teacher's future performance, even if it does contain real and potentially useful information (see Staiger & Rockoff 2010). The year-to-year correlation of teacher effect estimates has been found to range from 0.2 to 0.7, similar to objective performance measures in other jobs such as professional sports, insurance and security sales, and manual piece-rate production (McCaffrey et al. 2009). Nevertheless, given that many teacher labor contracts involve an "up or out" tenure decision fairly early in a teacher's career, usually after just two to four years, the apparent instability of teacher effect estimates raises concerns about using this information in personnel-related policy.

However, Staiger & Kane (2014) argue that year-to-year stability in annual performance is a misleading statistic. The impact of a retention decision, for instance, rests on the correlation between a single year's performance (or performance to date) and a teacher's remaining career performance. It is straightforward to show that the year-to-career correlation is just the square root of the year-to-year correlation so that a year-to-year correlation of 0.36 corresponds to a

[12]For example, if there are four fourth-grade teachers, each with one-quarter of the students, and a teacher with a value-added of -0.1 is replaced by a teacher with a value-added of 0.1, then the improvement in scores across cohorts should be $0.2/4 = 0.05$ standard deviations.

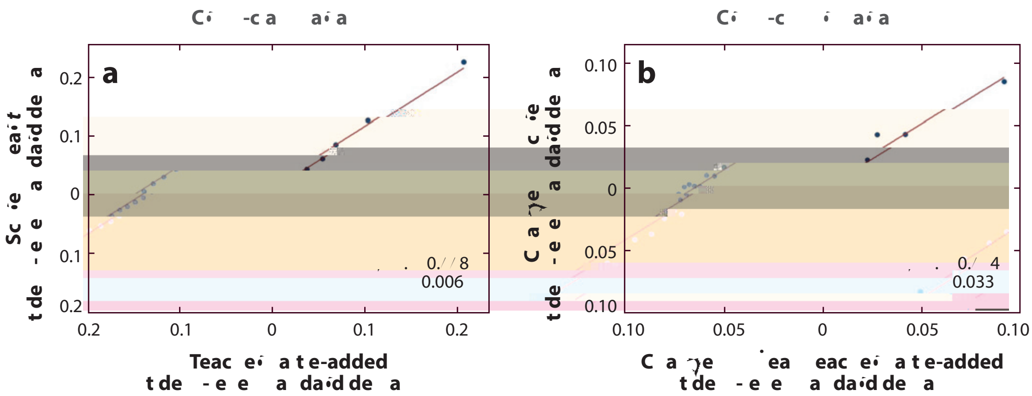

Figure 2

Quasi-experimental testing in Chetty et al. (2013a). (a) A binned scatter plot of test score residuals versus teacher value-added. (b) A plot of cross-cohort changes in mean test scores versus changes in mean teacher value-added at the school-grade level; these changes are also de-meaned by school year to eliminate secular time trends. Observations are divided into 20 equal-sized bins (vingtiles) based on the x-axis variable (value-added in a, change in mean value-added in b), and the mean of the y-axis variable (residual test score in a, change in mean score in b) within each group is plotted against the mean of the x-axis variable within each bin. The solid line shows the best linear fit estimated on the underlying micro data using ordinary least squares. The coefficients show the estimated slope of the best-fit line, with standard errors clustered at the school-cohort level reported in parentheses. Figure adapted from Chetty et al. (2013a, figures 2a and 4a).

year-to-career correlation of 0.6. Using data from several urban school districts that have six or more years of data on teachers' value-added, they estimate year-to-career correlations in the range of 0.65–0.8 for math and 0.55–0.7 for English. These imply, for example, that over three-quarters of teachers at the 25th percentile of one-year value-added have career value-added that is below average.

There are several ways in which noise in value-added measures might be reduced, most obviously by using multiple years of data, or generally multiple classrooms, to construct these measures. Lefgren & Sims (2012) combine teacher effects across subjects and find that the optimal weighted average of math and English value-added for elementary teachers in North Carolina substantially improved the ability of these measures to predict future teacher value-added in each individual subject. Alternatively, teacher effects could be combined with other sources of information. Mihaly et al. (2013) find that estimates of teachers' effectiveness are more stable when they incorporate classroom observations and student surveys but that these measures did not substantially improve the ability to predict teacher effects on test scores over what was possible using value-added estimates alone.[13]

ARE TEACHER EFFECTS STABLE ACROSS TIME AND CONTEXT?

Although most of the literature has assumed teacher effects to be fully persistent and fixed, recent evidence suggests that true teacher effects change over time and across different contexts, such as the school, grade, and subject being taught. Chetty et al. (2013a) and Goldhaber & Hansen (2013) estimate teacher effects with imperfect persistence and find that roughly half of the short-run persistence in teacher effects across classrooms in adjacent years is present among classrooms

[13]The in-class observation indicator they use is a teacher's scores on the Framework for Teaching evaluation rubric (Danielson 1996), averaged across four lessons. The student survey indicator is the previous year's class average response to questions on the Tripod Student Perceptions Survey (Ferguson 2009).

seven or more years apart. Similarly, Jackson (2013a) estimates teacher effects allowing for a school-specific match component. He finds that roughly half of the persistence in teacher effects observed across classrooms taught within the same school is present among classrooms taught in different schools. Evidence also suggests that there are subject- and grade-specific match components to teacher effects, with less persistence in teacher effects across classrooms taught in different grades (Kane & Staiger 2005) and across different subjects (Lefgren & Sims 2012, Condie et al. 2014).[14]

Changes in teacher effects across time and context have a number of important implications. First, estimates of teacher effects taken from a particular year or context will overstate the impact of that teacher in a different year or context—unless one uses estimation methods that allow for these changes, as suggested by Chetty et al. (2013a) and Lefgren & Sims (2012). Moreover, using such methods can yield improvement in the accuracy of teacher effect estimates, as discussed above. Finally, as Jackson (2013a) and Condie et al. (2014) suggest, the context-specific match component of teacher effects can be used to improve student perforchildren born when they are teenagers. Their results strongly support the idea that teacher effects on test scores have real economic content and capture, at least partially, a teacher's ability to raise students' human capital. Similarly, looking at high school teachers in North Carolina, Jackson (2013b) finds that teacher value-added in ninth-grade algebra and English predicts effects on dropping out, completing high schoolplel13Ue(n)1-1522.6ketinlpatinh itinditing tod tog thyear

TEACHER-RELATED POLICIES AIMED AT IMPROVING EDUCATIONAL PRODUCTION

Until recently, measures of teacher effectiveness such as those discussed above have played little role in teacher retention, evaluation, and pay decisions (Weisberg et al. 2009). However, as the evidence of wide variation in teacher effectiveness has grown more persuasive, many states and districts across the country have implemented teacher evaluation policies that incorporate value-added estimates, structured classroom observations, student perception surveys, and other methods to evaluate teachers. In the remainder of this section, we review recent findings in economics on using measures of teacher effectiveness for teacher selection, mentoring and feedback, and pay for performance.

Teacher Selection

As discussed above, differences in teacher effectiveness are large and persist over time. Although these differences are difficult to predict upon hiring based on teacher credentials, they can be predicted after observing a teacher's performance in the classroom. Accordingly, the evidence suggests that using measures of teacher effects for tenure or layoff decisions could improve the average effectiveness of the teacher workforce, as compared to the current practice of granting tenure as a matter of course to nearly all teachers and determining layoffs primarily based on seniority.

The potential for teacher selection has been illustrated in a variety of ways in a number of recent papers. Hanushek (2011) simulates the impact that removing the lowest-performing teachers would have on student test scores and earnings, making a range of plausible assumptions about the true variation across teachers, the amount of fade out, class size, and the relationship between achievement and earnings. He finds that replacing the bottom 5–10% of current teachers with teachers who had average effectiveness would raise average test scores

induction for one or two years, which included weekly meetings with a full-time mentor, monthly professional development sessions, opportunities to observe veteran teachers, and continuing evaluation of the teachers' own practices. Beginning teachers in control schools received any support normally offered to teachers by the school. There was no effect of the intervention in the first two years on students, measures of teaching practice, or teacher attrition. In the third year of coaching and evaluation, demonstrated improvements in observable teaching practices and student test scores that were approximately 5 percentile points (approximately 0.14σ) higher.

Taylor & Tyler (2012) study the effects of the evaluation system in Cincinnati Public Schools, in which teachers were evaluated based on specific criteria linked to higher achievement; teachers were observed in the classroom by peers and experts and received detailed feedback about areas in which they were deficient and how to improve.[21] Finally, the evaluation outcomes were linked to career development such that teachers who did not have strong evaluations had to undergo a year-long process of intensive assistance from a mentor that included another year of evaluation with more frequent observations. These evaluations are done periodically on a predetermined schedule, usually every five years, and the authors employ a quasi-experimental design, comparing the achievement of individual teachers' students before, during, and after the teacher's evaluation year. The authors find that students assigned to a teacher in a postevaluation year score approximately 0.1 standard deviations higher in math than similar students taught by the same teacher prior to evaluation. The magnitude of this estimate is notable given that the sample being studied comprised midcareer teachers, who many may have assumed could no longer acquire new skills. That the performance gains were sustained even after the evaluation year indicates that although incentive effects might have been important, these programs lead to real persistent increases in teacher skills. Although this may not be the only successful model of professional development, it is one that has been proven effective.

Pay for Performance

Public school teachers in the United States have traditionally been paid according to salary schedules based on years of experience and education level so that teacher pay is largely unresponsive to actual teacher performance (Podgursky & Springer 2007). In other contexts, worker effort and worker output are found to be higher when workers are paid for performance on the job (Foster & Rosenzweig 1994, Lazear 2000). If teaching is anything like other occupations, rewarding teachers for their performance may increase teacher effort and improve student outcomes.

Although performance pay is a promising idea, theoretically there are reasons why performance pay may be only weakly related to student achievement growth. First, we know from research on the estimation of teacher effects that student test scores are influenced by a variety of factors that are outside the control of the teacher. This problem can be reduced statistically by accounting for the influence of student attributes and family influences, much like the value-added approach discussed above. However, if these outside influences fluctuate over time in ways that are hard to predict, teachers may perceive a weak link between their effort and their pay, and accordingly not increase their effort. A second problem is that merit pay may not be effective at improving teacher performance if individual teachers do not know what to do to improve their teach88.4(a68aM8(are)Tm)-375.92wuurnanove 6t toed without value. This is a particularly acute problem in education, in which short-term measures such as test scores are only proxies for the development of human capital.[22] Thus, any well-designed pay-for-performance scheme must be based on outcomes that are a good measure of student learning and cannot be easily gamed. These potential problems underscore the importance of looking at the effects of pay for performance on unrewarded outcomes and looking for effects that persist over time.

Since the 1990s, pay for teacher performance has been adopted in many nations worldwide and in many districts in the United States, but it remains relatively uncommon.[23] Where there is a close correspondence between teacher effort and the rewarded performance, it is difficult to improve

Muralidharan & Sundararaman (2011) and Muralidharan (2012) analyze an experimental program in India that provided bonus payments to primary school teachers based on the average improvement of their students' test scores in independently administered learning assessments (with a mean bonus of 3% of annual pay). After two years, students in incentive schools performed significantly better than those in control schools, by 0.28 and 0.16 standard deviations in math and language tests, respectively. For students who completed five years of primary school under the program, test scores increased by 0.54 and 0.35 standard deviations in math and language, respectively, and also by 0.52 and 0.3 standard deviations in science and social studies, respectively, for which incentives were not provided.

We note that all these programs rewarded individual teachers for individual teacher outcomes and based rewards on average test scores rather than some proficiency cutoff. Taken together, these studies show that individual teacher incentive pay for average test score gains can lead to sizable improvements in student outcomes. Moreover, they demonstrate that this is true both in developing and in developed nations.

Evidence in the United States. There is substantial new evidence on pay for performance in the United States from randomized experiments, but with mixed results. Some studies show little effect of pay for performance, which has led some to speculate that performance pay cannot work in the United States. Other studies find positive impacts on incentivized outcomes, but only under particular design features or with negative spillovers onto nonincentivized outcomes. Thus, although it is possible that there is something different about teachers and students in the United States that renders teacher performance pay ineffective, design features of the programs may explain the differences in results.

Goodman & Turner (2013) and Fryer (2013) analyze a group incentive program in New York City. Under this program, a random sample of schools participated in a bonus pay scheme that involved team incentive pay at the school level linked to test score growth targets. The bonuses ranged from $1,500 to $3,000 per teacher (between 2.5% and 5% of the average teacher salary in New York City). The authors find that the bonus program had little impact on teacher effort, student performance in math and English, or classroom activities.

Goodman & Turner (2013) highlight the free-rider problem in how the program linked incentive pay to school-wide performance goals. They test for a free-rider problem by seeing if the effects are larger in smaller groups, in which the free-rider problem should be less severe, and find that this is indeed the case. This is also consistent with two studies using quasi-experimental methods to study teacher incentive programs in the United States as well as international evidence.[24]

Apart from the group incentive structure, this New York City program based rewards on a performance threshold, rather than rewarding general improvement.[25] Where teachers are responsible for average test scores, they have an incentive to improve the outcomes of all students. However, when teachers are responsible for reaching a performance threshold, they only have an

[24]Lavy (2002) and Muralidharan (2012) also find positive, but significantly smaller effects of group-based incentives in their respective settings. Sojourner et al. (2011) compare the effects of different kinds of pay-for-performance schemes in Minnesota and find that districts offering greater rewards for teacher-level goals experienced large gains in reading, whereas those offering rewards based on school-wide goals or subjective evaluations did not. Imberman & Lovenheim (2012) study the impact of a group-based performance pay system in Texas. Groups were defined at the subject-grade level, so the power of incentives directed at individual teachers varied both for the same teacher over time and for the same teacher in the same year across subjects and/or grade levels. They find robust evidence of weakened incentives when rewards are based on the collective performance of large groups of teachers, with ideal group sizes of three to four teachers.

incentive to expend effort on those students who can be pushed over this threshold (Neal & Schanzenbach 2010). If the performance threshold is too low, many schools can meet the standard by expending no additional effort, and if the performance threshold is too high, many schools will realize that they will not meet the standard, even if they expend additional effort, and will therefore chose not to do so. In the New York City program, almost 90% of schools earned awards, suggesting that the performance standard may have been too low to induce increased effort.

Another influential US-based finding of no effect of performance pay on test scores focuses on the POINT program in Tennessee. Under this system, middle school math teachers voluntarily participated in a controlled experiment and were randomly offered financial rewards for exhibiting "unusually large gains on standardized tests." Specifically, teachers could earn rewards of

teacher evaluation policies that go well beyond the existing evidence and urgently need answers

Boyd D, Lankford H, Loeb S, Wyckoff J. 2011. Teacher layoffs: an empirical illustration of seniority versus measures of effectiveness. *E . F* 6:439–54

Cascio EU, Staiger DO. 2012. . NBER Work. Pap. 18038

Chamberlain GE. 2013. Predictive effects of teachers and schools on test scores, college attendance, and earnings. *. A . A* 110:17176–82

Chetty R, Friedman JN, Rockoff JE. 2013a. . NBER Work. Pap. 19423

Chetty R, Friedman JN, Rockoff JE. 2013b. . NBER Work. Pap. 19424

Chingos MM, West MR. 2012. Do more effective teachers earn more outside the classroom? *E . F* 7:8–43

Clotfelter CT, Ladd HF, Vigdor JL. 2007. Teacher credentials and student achievement: longitudinal analysis with student fixed effects. *E . E .* 26:673–82

Clotfelter CT, Ladd HF, Vigdor JL. 2010. Teacher credentials and student achievement in high school: a cross-subject analysis with student fixed effects. *. .* 45:655–81

Condie S, Lefgren L, Sims D. 2014. Teacher heterogeneity, value added, and education policy. *E . E .* In press

Daly AJ, Moolenaar NM, Bolivar JM, Burke P. 2010. Relationships in reform: the role of teachers' social networks. *. E . A .* 48:359–91

Danielson C. 1996. *E : A F* . Alexandria, VA: ASCD

Dobbie W. 2011. *A* . Work. Pap., Harvard Univ., Cambridge, MA

Donaldson ML. 2009. . Rep., Cent. Am. Prog., Washington, DC

Ferguson R. 2009. *E E E* . Westwood, MA: Cambridge Ed.

Foster AD, Rosenzweig MR. 1994. A test for moral hazard in the labor market: contractual arrangements, effort, and health. *. E .* 76:213–27

Fryer RG. 2013. Teacher incentives and student achievement: evidence from New York City public schools. *. E .* 31:373–407

Fryer RG, Levitt SD, List J, Sadoff S. 2012. *E* . NBER Work. Pap. 18237

Glazerman S, Dolfin S, Bleeker M, Johnson A, Isenberg E, et al. 2008. . Rep., NCEE 2009-4034, US Dep. Educ., Natl. Cent. Educ. Eval. Reg. Assist., Inst. Educ. Sci., Washington, DC

Glazerman S, Isenberg E, Dolfin S, Bleeker M, Johnson A, et al. 2010. . Rep., NCEE 2010-4027, US Dep. Educ., Natl. Cent. Educ. Eval. Reg. Assist., Inst. Educ. Sci., Washington, DC

Goldhaber D, Chaplin D. 2012. *A : D* ? CEDR Work. Pap. 2011-5, Univ. Wash., Seattle

Goldhaber DD, Goldschmidt P, Tseng F. 2013. Teacher value-added at the high-school level: different models, different answers? *E . E . A .* 35:220–36

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

Goldhaber D, Theobald R. 2013. Managing the teacher workforce in austere times: the determinants and implications of teacher layoffs. *E . F* 8:494–527

Goldhaber D, Walch J. 2012. Strategic pay reform: a student outcomes-based evaluation of Denver's ProComp teacher pay initiative. *E . E .* 31:1067–83

Goodman SF, Turner LJ. 2013. The design of teacher incentive pay and educational outcomes: evidence from the New York City bonus program. *. E .* 31:409–20

Gordon R, Kane TJ, Staiger DO. 2006. . White Pap. 2006-01, Hamilton Proj., Washington, DC

Hamre BK, Pianta RC. 2010. Classroom environments and developmental processes: conceptualization and measurement. In *D*, ed. J Meece, J Eccles, pp. 25–41. New York: Routledge

Hanushek EA. 1971. Teacher characteristics and gains in student achievement: estimation using micro data. *A. E.* 61:280–88

Hanushek EA. 2011. The economic value of higher teacher quality. *E. E.* 30:466–79

Hanushek EA, Rivkin SG. 2010. Generalizations about using value-added measures of teacher quality. *A. E.* 100:267–71

Henry GT, Fortner CK, Bastian KC. 2012. The effects of experience and attrition for novice high-school science and mathematics teachers. 335:1118–21

Holcombe R, Jennings J, Koretz D. 2013. The roots of score inflation: an examination of opportunities in two states' tests. In *C, A E D*, ed. G Sunderman, pp. 163–89. Greenwich, CT: Inf. Age

Imberman SA, Lovenheim MF. 2012. . NBER Work. Pap. 18439

Ingersoll RM, Strong M. 2011. The impact of induction and mentoring programs for beginning teachers: a critical review of the research. *E.* 81:201–33

Isenberg E, Glazerman S, Bleeker M, Johnson A, Lugo-Gil J, et al. 2009. . Rep., NCEE 2009-4072, US Dep. Educ., Natl. Cent. Educ. Eval. Reg. Assist., Inst. Educ. Sci., Washington, DC

Jacob BA. 2005. Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago

Kane TJ, Staiger DO. 2005. [illegible]. Work. Pap., Dartmouth Coll., Hanover, NH
Kane TJ, Staiger DO. 2008.

Ost B. 2014. How do teachers improve? The relative importance of specific and general human capital. *A . E . . A. . E .* In press

Papay JP, Kraft MA. 2013. . Work. Pap., Harvard Univ., Cambridge, MA

Parsad B, Lewis L, Farris E, Greene B. 2001. *2000*. Rep., NCES 2001-88, US Dep. Educ., Natl. Cent. Educ. Stat., Washington, DC

Pianta RC. 2011. . Rep., Cent. Am. Prog., Washington, DC

Podgursky MJ, Springer MG. 2007. Teacher performance pay: a review. *. . A .* 26:909–50

Rivkin SG, Hanushek EA, Kain JF. 2005. Teachers, schools and academic achievement. *E* 73:417–58

Rockoff JE. 2004. The impact of individual teachers on student achievement: evidence from panel data. *A . E .* 94:247–52

Rockoff JE. 2008. *D ? E C* y. Unpublished manuscript, Columbia Bus. Sch.., New York

Rockoff JE, Jacob B, Kane TJ, Staiger DO. 2011. Can you recognize an effective teacher when you recruit one? *E . F* 6:43–74

Rockoff JE, Speroni C. 2010. Subjective and objective evaluations of teacher effectiveness. *A . E .* 100:261–66

Rockoff JE, Staiger DO, Kane TJ, Taylor ES. 2012. Information and employee evaluation: evidence from a randomized intervention in public schools. *A . E .* 102:3184–213

Rothstein J. 2009. Student sorting and bias in value-added estimation: selection on observables and unobservables. *E . F* 4:537–71

Rothstein J. 2010. Teacher quality in educational production: tracking, decay, and student achievement. *. . E .* 125:175–214

Rothstein J. 2012. . NBER Work. Pap. 18419

Slater H, Davies NM, Burgess S. 2012. Do teachers matter? Measuring the variation in teacher effectiveness in England. *. B . E .* 74:629–45

Sojourner A, West K, Mykerezi E. 2011. *? E -C* . Rep., Soc. Res. Educ. Eff., Evanston, IL

Annual Review of Economics

Volume 6, 2014

Contents

Errata

An online log of corrections to *Annual Review of Economics* articles may be found at http://www.annualreviews.org/errata/economics