CALDER

NATIONAL
CENTER for ANALYSIS of LONGITUDINAL DATA in EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

Urban Institute

A program of research by the Urban Institute with Duke University, Stanford University, University of Florida, University of Missouri-Columbia, University of Texas at Dallas, and University of Washington

BRIEF 9
MAY 2010

# Using Value-Added Measures of Teacher Quality

Eric A. Hanushek and Steven G. Rivkin

Extensive education research on the contribution of teachers to student achievement produces two generally accepted results. First, teacher quality varies substantially as measured by the value added to student achievement or future academic attainment or earnings. Second, variables often used to determine entry into the profession and salaries—including postgraduate schooling, experience, and licensing examination scores—appear to explain little of the variation in teacher quality so measured. (Early experience is the exception.) Together, these findings underscore explicitly that observed teacher characteristics do not represent teacher quality.

*Value-added research strongly supports the presence of substantial differences in teacher effectiveness, even within schools.*

From the earliest work on education productions (Coleman et al. 1966), interpretations of research on teachers often confused the effects of specific teacher characteristics with the general contribution of teachers. Research over four decades has consistently found that the most common indicators of quality differences are not closely related to achievement gains, leading some to question whether teacher quality really matters (see review in Hanushek and Rivkin 2006).

Education production function research on the measurement of teacher value added to student achievement has shifted from a research framework that focuses on the link between student outcomes and specific teacher characteristics to a framework that uses a less parametric approach to identify teacher contributions to learning. Using longitudinal administrative databases, some covering all teachers within a state, value-added research strongly supports the presence of substantial differences in teacher effectiveness, even within schools. Although this approach circumvents the need to identify specific teacher characteristics related to quality, it introduces additional complications and has sparked an active debate on the measurement and subsequent policy use of estimated teacher value added.

## BASIC ANALYTICAL FRAMEWORK AND FINDINGS

The precise method of attributing differences in classroom achievement to teachers is the subject of considerable discussion and analysis. We begin by briefly outlining the general analytical framework that forms the basis of much of the work in this area and then describe the range of results from recent efforts to measure the variance of teacher effectiveness.

Analyses of teacher value added typically begin with an education production function:

$$A_g = \theta A_{g-1} + \tau_j + \mathbf{S}\varphi + \mathbf{X}\gamma + \varepsilon$$

*Having a teacher at the 25th percentile of the quality distribution versus the 75th percentile would move a student at the middle of the achievement distribution to the 59th percentile.*

where $A_g$ is the achievement of student $i$ in grade $g$ (the subscript i is suppressed throughout), $A_{g-1}$ is the prior-year student achievement in grade $g$-1, $S$ is a vector of school and peer factors, $X$ is a vector of family and neighborhood inputs, $\theta$,$\varphi$, and $\gamma$ are unknown parameters, $\varepsilon$ is a stochastic term representing unmeasured influences, and $\tau_j$ is a teacher fixed effect that provides a measure of teacher value added for teacher $j$. (Alternative estimation forms, largely restricting $\theta$, have pluses and minuses but are currently less frequently employed; see Rivkin 2005).

Table 1 summarizes existing estimates of the standard deviation of $\tau_j$ expressed in units of student achievement (normalized to a standard deviation of one). Covering a range of schooling environments across the United States, these studies produce fairly similar estimates of the variance in teacher value added: the average standard deviation for reading is 0.13 and for math is 0.17, and the distributions for both are fairly tight. Note also that except for Kane and Staiger (2008), these estimates rely on just within-school variation in value added, ignoring the surprisingly small between-school component (not typically considered because of potential sorting, testing, and other interpretative problems).

**Table 1. Estimated Standard Deviation of Teacher Effectiveness Measured in Standard Deviations of Student Achievement**

| | | Teacher Effectiveness | |
|---|---|---|---|
| Study | Location | Reading | Math |
| Rockoff (2004) | New Jersey | 0.10 | 0.11 |
| Nye, Konstantopoulos, and Hedges (2004) | Tennessee | 0.26 | 0.36 |
| Rivkin, Hanushek, and Kain (2005) | Texas | 0.10 | 0.11 |
| Aaronson, Barrow, and Sander (2007) | Chicago | — | 0.13 |
| Kane, Rockoff, and Staiger (2008) | New York City | 0.08 | 0.11 |
| Jacob and Lefgren (2008) | Undisclosed city | 0.12 | 0.26 |
| Kane and Staiger (2008) | Los Angeles | 0.18 | 0.22 |
| Koedel and Betts (2009) | San Diego | — | 0.23 |
| Jesse Rothstein (2010) | North Carolina | 0.11 | 0.15 |
| Hanushek and Rivkin (2010) | Undisclosed city | — | 0.11 |

Notes: All estimates indicate the standard deviation of teacher effectiveness in terms of student achievement standardized to mean 0 and variance 1. All variances are corrected for test measurement error and, except Kane and Staiger (2008), are estimated within school-by-year or within school-by-grade-by-year.

The magnitudes of these estimates support the beliefs that teacher quality is an important determinant of school quality and achievement. For example, the math results imply that having a teacher at the 25th percentile of the quality distribution versus the 75th percentile would mean a difference in learning gains of roughly 0.2 standard deviations in a single year. This would move a student at the middle of the achievement distribution to the 59th percentile. The magnitude of such an effect is large relative to both typical measures of black-white or income achievement gaps (0.7–1.0 standard deviations) and methodologically compelling estimates of the effects of a 10-student reduction in class size (0.1–0.3 standard deviations).

## METHODOLOGICAL CONCERNS

Of course, the value of these estimates hinges upon a number of factors including the relevance of the test instrument, the consistency of the estimator, and the persistence of teacher quality effects. A growing body of work considers these issues (Jacob, Lefgren, and Sims 2008; Kane and Staiger 2008; Ishii and Rivkin 2009; Rothstein 2010). We focus our discussion on test measurement and the empirical methods used to estimate $\tau_j$.

The testing questions have several components. One fundamental question—do these tests measure skills that are important or valuable?—appears well answered, as research demonstrates that standardized test scores relate closely to school attainment, earnings, and aggregate economic outcomes (Murnane, Willett, and Levy 1995; Hanushek and Woessmann 2008). The one caveat is that this body of research is based on low-stakes tests that do not affect teachers or schools. The link between test scores and high-stakes tests might be weaker if such tests lead to more narrow teaching, more cheating, and so on.

Another testing issue involves measurement error, a complication that takes on added importance in residual-based estimates of the variance of teacher quality. No achievement test completely and accurately measures true student knowledge. The selection of specific questions,

random events surrounding testing situations, familiarity with the tests, and other factors can lead measured scores to differ from true, underlying student knowledge, and these test errors will propagate into errors in estimates of value added for teachers. All but one variance estimate in table 1 is actually adjusted for measurement error, and the adjustment substantially reduces the estimated variance in teacher quality. Across the six studies that provide sufficient data, the variance in measurement error is only slightly smaller than the variance in true effectiveness when estimating on a school-year basis.

A final set of measurement issues relates to the details of test measurement: Do available tests emphasize a particular range (typically basic skills) more than others? Is there a ceiling on test performance? Is there an interval scale for test scores? The implication of each is that the estimated value added of teachers appears to depend specifically on test details. Although evidence suggests that these matters deserve attention, such complications do not appear to threaten the basic result that teacher quality varies substantially.

*No achievement test completely and accurately measures true student knowledge.*

Separate issues about value-added estimation relate to whether omitted variables lead to biased estimates of $\tau_j$. Specifically, if the empirical model fails to account for student differences that affect school choice, then estimates of teacher effects and the aggregate variance could be biased. These are particularly complex issues, given that both parents and school personnel exercise choices (see Hanushek, Kain, and Rivkin 2004a, 2004b). These issues have been a matter of concern for a long time (e.g., Hanushek 1992); as a result, all but one estimate in table 1 focuses solely on within-school differences in teacher performance.

More recent formalization and empirical analysis by Rothstein (2010) has emphasized classroom sorting and selection. In this work, the possibility for nonrandom classroom assignment to yield biased estimates of teacher value added is analyzed using North Carolina achievement data. For the models presented in table 1, the analysis suggests that the standard deviation of bias could be around 20 percent statewide and possibly much larger in schools that track on the basis of prior achievement.

A compelling part of Rothstein's analysis is the development of falsification tests, where future teachers are shown to significantly affect current achievement. Although this difference could be driven in part by subsequent-year classroom placement based on current achievement, the analysis suggests the presence of additional unobserved differences.

In related work, Hanushek and Rivkin (2010) use alternative, albeit imperfect, methods for judging which schools systematically sort students in a large Texas district. In the sorted samples, where random classroom assignment is rejected, this falsification test performs like that in North Carolina. But this is not the case in the remaining unsorted sample, where random assignment is not rejected. Kane and Staiger's (2008) alternative approach that uses estimates from a random assignment of teachers to classrooms finds little bias in traditional estimation, although the possible uniqueness of the sample and the limitations of the specification test suggest care in interpretation of the results.

The variance estimates of Rivkin, Hanushek, and Kain (2005) rely on a different estimation approach that guards against such sorting but likely produces downward-biased estimates of the variance in teacher quality. As table 1 shows, these estimates tend to be below the others in the table, with the difference across studies in the range of the bias estimated by Rothstein (2010). Therefore, although the impact of any classroom sorting on unobservables remains an important and unresolved question, the finding of substantial variation in teacher quality appears robust to such sorting.

## THE POLICY USES OF TEACHER VALUE ADDED

The attention to estimation of value-added models clearly results from the potential policy uses of such estimation. Collectively, there appears little doubt that there are significant differences in teacher effectiveness —and that actions to improve the quality of teachers could dramatically affect U.S. achievement. For example, Hanushek (2009) uses estimates of variations in the range of table 1 and shows that eliminating 6–10 percent of the worst teachers could have strong impacts on student achievement, even if these teachers were replaced permanently with just average teachers.

The bigger issues with value-added estimates of teacher effectiveness concern their use in personnel compensation, employment, promotion, or assignment decisions. The possibility of introducing performance pay based on value-added estimates motivates much of the prior analysis of the properties of these estimates, but movement in this direction has so far been limited (Podgursky and Springer 2007). Despite the strength of the research findings, concerns about accuracy, fairness, and potential adverse effects of incentives based on limited outcomes raise worries about using value-added estimates in education staffing and policy decisions. Many possible drawbacks relate to the measurement and estimation issues discussed above, but there are also concerns about incentives to cheat, adopt teaching methods that teach narrowly to tests, and ignore nontested subjects.

*The key policy question is whether value-added measures, despite shortcomings, can provide valuable information to improve personnel decisions.*

Although researchers can mitigate the effects of sampling error on estimates of teacher quality, such error would inevitably lead some successful teachers to receive low ratings and some unsuccessful teachers to receive high ratings. The measurement error issues largely go away if teachers are observed over multiple years and with large numbers of children (McCaffrey et al. 2009). However, relying on multiple years of data eliminates new teachers from any system and dampens the strength of incentives, as job performance in the current year would only partially determine the measure of effectiveness.

In terms of fairness, any failure to account for sorting on unobservable characteristics could penalize teachers given unobservably more difficult classrooms and reward teachers given unobservably less difficult classrooms. This could discourage educationally beneficial decisions including the assignment of more difficult or disruptive students to higher-quality teachers. This potential drawback can, however, be mitigated by combining subjective supervisor or peer evaluations with objective value-added estimates, since principals could place the estimates in context and appear able to judge differences in effectiveness at least at the tails of the distribution (Jacob and Lefgren 2008).

Finally, concentrating on within-school variation may not be appropriate for policy. The within-school focus, taken because of the difficulty accounting for differences among schools, raises concerns for performance evaluation; some schools may have much better teachers on average than others, and it would be important to recognize such differences.

All in all, cataloguing the potential imperfections of value-added measures is simple, but so is cataloging the imperfections of the current system with limited performance incentives and inadequate evaluations of teachers and administrators. Potential problems certainly suggest that statistical estimates of quality based on student achievement in reading and mathematics should not constitute the sole component of any evaluation system. Even so, the key policy question is whether value-added measures, despite shortcomings, can provide valuable information to improve personnel decisions that currently rely on limited information about teacher effectiveness and often provide weak performance incentives to teachers and administrators. The case for objective measures is likely strongest in urban or rural areas where there is more limited competition among public and private schools. In such places, a hybrid approach to evaluation in which value-added measures constitute one of several components may have great promise.

## REFERENCES


Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of Labor Economics* 25(1): 95–135.

Coleman, James S., Ernest Q. Campbell, Carol J. Hobson, James McPartland, Alexander M. Mood, Frederic D. Weinfeld, and Robert L. York. 1966. *Equality of Educational Opportunity*. Washington, DC: U.S. Government Printing Office.

Hanushek, Eric A. 1992. "The Trade-Off between Child Quantity and Quality." *Journal of Political Economy* 100(1): 84–117.

———. 2009. "Teacher Deselection." In *Creating a New Teaching Profession*, edited by Dan Goldhaber and Jane Hannaway (165–80). Washington, DC: Urban Institute Press.

Hanushek, Eric A., and Steven G. Rivkin. 2006. "Teacher Quality." In *Handbook of the Economics of Education,* edited by Eric A. Hanushek and Finis Welch (1051–78). Amsterdam: North Holland.

———. 2010. "Constrained Job Matching: Does Teacher Job Search Harm Disadvantaged Urban Schools?" Working paper 15816. Cambridge, MA: National Bureau of Economic Research.

Hanushek, Eric A., and Ludger Woessmann. 2008. "The Role of Cognitive Skills in Economic Development." *Journal of Economic Literature* 46(3): 607–68.

Hanushek, Eric A., John F. Kain, and Steven G. Rivkin. 2004a. "Disruption versus Tiebout Improvement: The Costs and Benefits of Switching Schools." *Journal of Public Economics* 88(9/10): 1721–46.

———. 2004b. "Why Public Schools Lose Teachers." *Journal of Human Resources* 39(2): 326–54.

Ishii, Jun, and Steven G. Rivkin. 2009. "Impediments to the Estimation of Teacher Value Added." *Education Finance and Policy* 4(4): 520–36.

Jacob, Brian A., and Lars Lefgren. 2008. "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education." *Journal of Labor Economics* 26(1): 101–36.

Jacob, Brian A., Lars Lefgren, and David P. Sims. 2008. "The Persistence of Teacher-Induced Learning Gains." Working Paper 14065. Cambridge, MA: National Bureau of Economic Research.

Kane, Thomas J., and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." Working Paper 14607. Cambridge, MA: National Bureau of Economic Research.

Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger. 2008. "What Does Certification Tell Us about Teacher Effectiveness? Evidence from New York City." *Economics of Education Review* 27(6): 615–31.

Koedel, Cory, and Julian R. Betts. 2009. "Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique." Working Paper 09-02. Columbia: University of Missouri.

McCaffrey, Daniel F., Tim R. Sass, J. R. Lockwood, and Kata Mihaly. 2009. "The Intertemporal Variability of Teacher Effect Estimates." *Education Finance and Policy* 4(4): 572–606.

Murnane, Richard J., John B. Willett, and Frank Levy. 1995. "The Growing Importance of Cognitive Skills in Wage Determination." *Review of Economics and Statistics* 77(2): 251–66.

Nye, Barbara, Spyros Konstantopoulos, and Larry V. Hedges. 2004. "How Large Are Teacher Effects?" *Educational Evaluation and Policy Analysis* 26(3): 237–57.

Podgursky, Michael J., and Matthew G. Springer. 2007. "Teacher Performance Pay: A Review." *Journal of Policy Analysis and Management* 26(4): 909–49.

Rivkin, Steven G. 2005. "Cumulative Nature of Learning and Specification Bias in Education Research." Mimeo, Amherst College.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73(2): 417–58.

Rockoff, Jonah E. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *American Economic Review* 94(2):247–52.

Rothstein, Jesse. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics* 125(1): 175–214.

This paper was presented at the annual meeting of the American Economic Association held in Atlanta, Georgia, in January 2010. It, along with other papers presented in the session "Implicit Measurement of Teacher Quality," is published in *The American Economic Review* 100(2), May 2010.

## ABOUT THE AUTHORS

Eric A. Hanushek is the Paul and Jean Hanna Senior Fellow at the Hoover Institution of Stanford University. He has been a leader in the development of economic analysis of educational issues, and his work on efficiency, resource use, and economic outcomes of schools has frequently entered into the design of both national and international educational policy. His analysis measuring teacher quality through student achievement forms the basis for current research into the value added of teachers and schools. He has produced 15 books along with numerous widely cited articles in professional journals. Dr. Hanushek is also chairman of the executive committee for the Texas Schools Project at the University of Texas at Dallas, and a member of the CALDER management team, leading the CALDER Texas work. He has served as deputy director of the Congressional Budget Office. His full C.V. is available at www.caldercenter.org.

Steven G. Rivkin is professor of economics and chair of the department of economics at Amherst College, associate director of research with the Texas Schools Project at the University of Texas at Dallas, and a member of the CALDER Texas research team. He is also a member of the Town of Amherst (MA) School Committee. His research centers on the economics and sociology of education as it relates to policy discussions Dr. Rivkin has written extensively on teacher quality, teacher labor markets, class size effects, school spending, and school desegregation; he has also written studies on the effectiveness of charter schools and parental responsiveness to charter school quality, special education, student mobility, peer influences, and the effects of air pollution on absenteeism. His full C.V. is available at www.caldercenter.org.

THE URBAN INSTITUTE
2100 M Street, N.W.
Washington, D.C. 20037

Phone: 202-833-7200
Fax: 202-467-5775
http://www.urban.org



National Center for Analysis of Longitudinal Data in Education Research

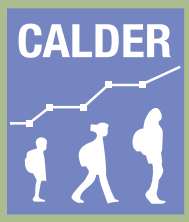
IN THIS ISSUE

**Using Value-Added Measures of Teacher Quality**

To download this document, visit our web site, http://www.urban.org.

For media inquiries, please contact paffairs@urban.org.

National Center for Analysis of Longitudinal Data in Education Research




*This research is part of the activities of the National Center for the Analysis of Longitudinal Data in Education Research (CALDER). CALDER is supported by Grant R305A060018 to the Urban Institute from the Institute of Education Sciences, U.S. Department of Education. More information on CALDER is available at http://www.caldercenter.org.*

*This paper was presented at the annual meeting of the American Economic Association held in Atlanta, Georgia, in January 2010. It, along with other papers presented in the session "Implicit Measurement of Teacher Quality," is published in* The American Economic Review *100(2), May 2010.*