American Economic Review: Papers & Proceedings 100 (May 2010): 267–271
http://www.aeaweb.org/articles.php?doi=10.1257/aer.100.2.267

Generalizations about Using Value-Added Measures of Teacher Quality

By Eric A. Hanushek and Steven G. Rivkin*

The extensive investigation of the contribution of teachers to student achievement produces two generally accepted results. First, there is substantial variation in teacher quality as measured by the value added to achievement or future academic attainment or earnings. Second, variables often used to determine entry into the profession and salaries, including post-graduate schooling, experience, and licensing examination scores, appear to explain little of the variation in teacher quality so measured, with the exception of early experience. Together these findings underscore explicitly that observed teacher characteristics do not represent teacher quality.

From the earliest work on education productions (James S. Coleman et al. 1966), interpretations of research on teachers often confused the effects of specific teacher characteristics with the overall contribution of teachers. The consistent finding over four decades has been that the most commonly used indicators of quality differences are not closely related to achievement gain, leading some to question whether teacher quality really matters (see the review in Eric A. Hanushek and Steven G. Rivkin 2006).

Education production function research on the measurement of teacher value added to student achievement represents a shift from a research design that focuses on the link between student outcomes and specific teacher characteristics to a research framework that uses a less parametric approach to identify overall teacher contributions to learning. Using administrative databases, some covering all of the teachers in a state, such research provides strong support for the existence of substantial differences in teacher effectiveness, even within schools. Although this approach circumvents the need to identify specific teacher characteristics related to quality, the less parametric approach introduces additional complications and has sparked an active debate on the measurement and subsequent policy use of estimated teacher value added.

* Hanushek: Hoover Institution, Stanford University, Stanford, CA 94305 (e-mail: hanushek@stanford.edu); Rivkin: Department of Economics, Amherst College, Amherst, MA 01002 (e-mail sgrivkin@amherst.edu).

I. 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:

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

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 $g - 1$, $\mathbf{S}$ is a vector of school and peer factors, $\mathbf{X}$ is a vector of family and neighborhood inputs, θ, φ, and γ are unknown parameters, ε is a stochastic term representing unmeasured influences, and τ_j is a teacher fixed effect that provides a measure of teacher value added for teacher j. (Alternative estimation forms, largely restricting θ, have pluses and minuses but are currently less frequently employed; see Rivkin 2006.)

Table 1 summarizes existing estimates of the standard deviation of τ_j expressed in units of student achievement (normalized to a standard deviation of one). Although 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 0.11 and for math is 0.15, and the distributions for both are fairly tight. Note also 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).

The magnitudes of these estimates support the belief that teacher quality is an important determinant of school quality and achievement. For example, the math results imply that having a teacher at the twenty-fifth percentile as compared to the seventy-fifth percentile of the quality distribution 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 fifty-eighth percentile. The magnitude of such an effect is large both relative to typical measures of black-white or income achievement gaps of 0.7–1 standard deviation and compared to methodologically compelling estimates of the effects of a ten student reduction in class size of 0.1–0.3 standard deviations.

II. Methodological Concerns

Of course the value of these estimates hinges upon a number of factors, including the relevance of the test instrument, consistency of the estimator, and the persistence of teacher quality effects. A growing body of work considers these issues (e.g. Jacob, Lefgren, and David Sims 2008, Kane and Staiger 2008, Jun Ishii and Rivkin 2009, and Rothstein 2010). We focus our discussion on test measurement and the empirical methods used to estimate τ_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 are closely related to school attainment, earnings, and aggregate economic outcomes (e.g., Richard J. Murnane, John B. Willett, and Frank Levy 1995 and Hanushek and Ludger 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 economic outcomes and high-stakes tests might be weaker if such tests lead to more narrow teaching, more cheating, etc.

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 of the variance estimates 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 estimation is done 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. Yet, although existing evidence suggests that these matters deserve attention, such complications do not appear to threaten the basic result that there is substantial variation in teacher quality.

A separate set of issues about value added estimation relates to whether omitted variables lead to biased estimates of τ_j. Specifically, if the empirical model fails to account for student differences that affect school choice, 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 (c.f. Hanushek, Kain, and Rivkin 2004a, b). These issues have been a matter of concern for a long time (e.g., Hanushek 1992), and as a result, all but one of the estimates in Table 1 focuses solely on within-school differences in teacher performance.

More recent formalization and empirical analysis by Rothstein (2010) emphasizes classroom sorting and selection. In this work, the possibility that nonrandom classroom assignment yields biased estimates of teacher value added is analyzed with North Carolina achievement data. For the models presented in Table 1, the analysis suggests that the standard deviation of bias could be on the order of 20 percent in North Carolina as a whole and possibly much

Table 1—Estimated Standard Deviation of Teacher Effectiveness Measured in Terms of Standard Deviations of Student Achievement

Study	Location	Teacher effectiveness (SD)	
		Reading	Math
Jonah E. Rockoff (2004)	New Jersey	0.10	0.11
Barbara Nye, Spyros Konstantopoulos and Larry V. Hedges (2004)	Tennessee	0.26	0.36
Rivkin, Hanushek and John F. Kain (2005)	Texas	0.10	0.11
Daniel Aaronson, Lisa Barrow and William Sander (2007)	Chicago		0.13
Thomas J. Kane, Jonah E. Rockoff and Douglas O. Staiger (2008)	New York City	0.08	0.11
Brian A. Jacob and Lars Lefgren (2008)	Undisclosed city	0.12	0.26
Kane and Staiger (2008)	Los Angeles	0.18	0.22
Cory Koedel and Julian R. 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 zero and variance one. 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.

larger in schools that track on the basis of prior achievement.

A compelling part of the analysis in Rothstein (2010) is the development of falsification tests, where future teachers are shown to have significant effects on current achievement. Although this 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. An alternative approach of Kane and Staiger (2008) of using 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.

Interestingly, 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 do tend to be below the others in the table, with the difference across studies being in the range of the bias estimated by Rothstein (2010). Thus although the impact of any classroom sorting on unobservables remains an important and unresolved question, the finding of substantial variation in teacher quality appears to be robust to such sorting.

III. Policy Uses of Value-Added Estimates of Teacher Effectiveness

The attention to estimation of value-added models clearly results from the potential policy uses of such estimation. At the aggregate level, there appears little doubt that there are significant differences in teacher effectiveness—and that actions to improve the quality of teachers could have a dramatic effect on US 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 a dramatic impact on student achievement even if these were replaced (permanently) with just average teachers.

The bigger set of issues, however, relates to the use of teacher value-added estimates in 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 (Michael J. Podgursky and Matthew G. Springer 2007). Despite the strength of the research findings, concerns about accuracy, fairness, and potential adverse effects of incentives based on

a limited set of outcomes raise worries about the use of value added estimates in education personnel and policy decisions. Many of the possible drawbacks are related 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 non-tested subjects.

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 (Daniel F. 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 would potentially 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 to be able to judge differences in effectiveness at least at the tails of the distribution (Jacob and Lefgren 2008).

Finally, concentration 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, since 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. Nonetheless, the key policy question is whether the value of even flawed value added measures could advance the current system of personnel decisions that relies on limited information about teacher effectiveness and often provides weak performance incentives to teachers and administrators. The case in support of objective measures is likely to be 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 a number of 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., Ernst 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.

Hanushek, Eric A. 2009. "Teacher Deselection." In *Creating a New Teaching Profession*, ed. 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*. Vol. 1, ed. Eric A. Hanushek and Finis Welch, 1051–78. Amsterdam: North-Holland.

Hanushek, Eric A., and Steven G. Rivkin. 2010. "Constrained Job Matching: Does Teacher Job Search Harm Disadvantaged Urban Schools?" National Bureau of Economic Research Working Paper 15816.

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.

Hanushek, Eric A., John F. Kain, and Steven G. Rivkin. 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 Sims. 2008. "The Persistence of Teacher-Induced Learning Gains." National Bureau of Economic Research Working Paper 14065.

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.

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

Koedel, Cory, and Julian Betts. 2009. "Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique." University of Missouri Department of Economics Working Paper 0902.

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. 2006. "Cumulative Nature of Learning and Specification Bias in Education Research." Unpublished.

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.