Generalizations about Using Value-Added Measures of Teacher Quality

by Eric A. Hanushek and Steven G. Rivkin

Paper presented at the annual meetings of the

American Economic Association
Atlanta, Georgia

January 2010

Generalizations about Using Value-Added Measures of Teacher Quality

by Eric A. Hanushek and Steven G. Rivkin[1]

The extensive investigation of the role of teachers in student performance produces two generally accepted results. First, there is substantial variation in teacher quality as measured by the contribution 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, even those often used in salary determination or as hurdles to enter the profession, do not represent teacher quality.

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 overall contribution of teachers. The consistent finding over four decades has been that the most commonly used indicators of quality differences – teacher education, teacher experience, teacher certification, and the like – are not closely related to achievement (Hanushek and Rivkin (2006)). With a few exceptions such as the typical gains to the first years of experience, none of the measures used in either research or policy has been shown to be consistently related to effectiveness in the classroom, leading some to question whether teacher quality really matters.

Recent 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 of teachers to a research framework that uses a less parametric approach to identify overall teacher contributions to learning. Using administrative data bases, some covering all of the teachers in a state, such

[1] Hoover Institution, Stanford University, Stanford, CA 94305 (hanushek@stanford.edu); Department of Economics, Amherst College, Amherst, MA 01002 (sgrivkin@amherst.edu).

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 characteristics of teachers related to teacher quality, the less parametric approach introduces additional complications into research on teacher quality and has sparked an active debate on the measurement of 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 now briefly outline 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 some form of this general education production function:

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

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, *S* is a vector of school factors including peer composition, *X* is a vector of family and neighborhood inputs, 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 (2005)).

Table 1 provides a summary of estimates of the standard deviation of τ_j expressed in units of student achievement (normalized to a standard deviation of one). Although these studies cover a range of schooling environments across the US, they produce fairly similar estimates of the variance in teacher value added: the average standard deviation for reading is 0.11 and that

for math is 0.15, and the distributions for both are fairly tight. Note also that these estimates rely on just within-school variation in value-added, ignoring the surprisingly small between-school component because of potential sorting, testing, and other interpretative problems.

The magnitudes of these estimates support the beliefs that differences in teacher quality make substantial contributions to overall achievement differences and that teacher quality differences are of primary importance in the determination of school quality. For example, the math results imply that having a teacher at the 25th percentile as compared to the 75th 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 58th 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.

Methodological Concerns with Estimated Value-Added

Of course the value and policy relevance of these estimates hinges upon a number of factors including the relevance of the test instrument, the success of the empirical methods in the identification of teacher contributions to learning and the persistence of teacher quality effects, and a growing body of work considers these issues (e.g., Ishii and Rivkin (2009), 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. The common tests are designed to cover a given domain of knowledge, but they generally never consider any

external validation – whether for college choices or for labor market outcomes. Fortunately, other research demonstrates that the measured skills are closely related to school attainment, earnings, and aggregate economic outcomes (e.g., Murnane, Willett, and Levy (1995) and Hanushek and Woessmann (2008)). The one caveat is that this body of research is based on low-stakes tests, and the link between test scores 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 in 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 ability, and these test errors will in turn lead to 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 (usually thought of as basic skills) more than others? Is there ceiling on test performance? Is there an interval scale for test scores? The implication of each is that the estimated value-added of teachers may depend specifically on the range of student performance being considered. Unfortunately, the answer to this set of questions appears to depend crucially on the specific test being analyzed, and even then there are many open questions in existing research – making this area important but largely unanswered. Existing

evidence suggests that these matters deserve attention but that such complications do not 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, often dealing with selection effects, leads to bias in the estimated τ_j. Specifically, if the selection into the schools or classroom – either because of student and parent choices or teacher and principal choices – leads to specific kinds of classroom and teacher matches, the individual teacher and the aggregate variance estimation might be biased. These are particularly complex issues, given at the outset that both parents and school personnel are known to be concerned about school choice (c.f. Hanushek, Kain, and Rivkin (2004a, (2004b)). 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 relies upon within-school variance in quality and eliminates all of the between-school variations in student performance.

More recent formalization and empirical analysis by Rothstein (2009, (2010) has emphasized classroom sorting and selection. In this work, the possibility for nonrandom classroom assignment leading to biased estimates of teacher value-added is analyzed with the North Carolina achievement data. For the models presented in Table 1, the North Carolina analysis suggests that the standard deviation of bias could be on the order of one-fifth of the total estimates. These estimates are average effects across schools with and without unmeasured sorting, so the effects could be noticeably larger in sorted schools if there is also a significant population of unsorted schools.

A powerful part of the analysis in Rothstein (2010) is the development of falsification tests, where future teachers (who have not been in class with any of the students) are shown to have significant effects on current achievement. Although this could be part of classroom

placement on observable achievement, the analysis suggests that the effects go beyond that. 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 does not hold 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 limitations of this type of specification test suggest care in interpretation of the results.

Interestingly, the estimates of Rivkin, Hanushek, and Kain (2005) rely on a different estimation approach that guards against sorting but yields lower bounds estimates of the variance in teacher effectiveness. From Table 1, these estimates do tend to be below the others in the table, with the difference across studies being in the range of bias by the Rothstein (2010) estimates. 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 be robust to such sorting.

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 dramatic effects on 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 dramatic impacts on student achievement even if these were just replaced (permanently) with average teachers.

The bigger set of issues, however, relates to the use of individual teacher estimates of value-added in teacher compensation, employment, assignment, or promotion decisions. 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. Sole reliance on a single approach is unlikely to work well, implying that balancing and combining the alternatives is the central issue. The possibility of introducing more direct 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)).

Concerns about the precision and fairness of measures of teacher value-added and potential adverse effects of incentives based on a limited set of outcomes raise worries about any expanded use of value added estimates in education personnel and policy decisions, despite the strength of the research findings. Many of the possible drawbacks are related to the test measurement and estimation issues discussed above, but there are also concerns about increased incentives to cheat on tests, 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 the variance in 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 individual teachers are observed over multiple years and with substantial numbers of children (McCaffrey, Sass, Lockwood, and Mihaly (2009)). However,

relying on multiple years of data eliminates new teachers from any system and dampens the power of incentives.

In terms of fairness, any failure to account for sorting on unobservables 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. Moreover, all of the estimates presented in Table 1 were based upon within-school variation due to the acknowledged difficulty of accounting for between school differences in other school and community factors. The within-school aspect presents some concerns for any performance evaluation, because some schools may have much better teachers on average than others, and it would be important to recognize such differences. Some of these issues can, nonetheless, be mitigated by including subjective supervisor or peer evaluations with objective value-added estimates, since principals seem to be able to judge differences in effectiveness at least at the tails of the distribution (Jacob and Lefgren (2008)).

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, such objective estimates also contain valuable information that could advance the current system that relies on limited information about teacher effectiveness and provides few performance incentives, particularly in urban or rural areas where the competition among schools is stunted.

References

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. "Teachers and Student Achievement in the Chicago Public High Schools." Journal of Labor Economics 25,no.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, D.C.: U.S. Government Printing Office.

Hanushek, Eric A. 1992. "The trade-off between child quantity and quality." Journal of Political Economy 100,no.1 (February):84-117.

———. 2009. "Teacher deselection." In Creating a new teaching profession, edited by Dan Goldhaber and Jane Hannaway. Washington, DC: Urban Institute Press:165-180.

Hanushek, Eric A., John F. Kain, and Steve G. Rivkin. 2004a. "Disruption versus Tiebout improvement: The costs and benefits of switching schools." Journal of Public Economics Vol 88/9-10:1721-1746.

———. 2004b. "Why public schools lose teachers." Journal of Human Resources 39,no.2:326-354.

Hanushek, Eric A., and Steve G. Rivkin. 2010. "Constrained Job Matching: Does Teacher Job Search Harm Disadvantaged Urban Schools?" mimeo, Hoover Institution, Stanford University (

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. Amsterdam: North Holland:1051-1078.

Hanushek, Eric A., and Ludger Woessmann. 2008. "The role of cognitive skills in economic development." Journal of Economic Literature 46,no.3 (September):607-668.

Ishii, Jun, and Steven G. Rivkin. 2009. "Impediments to the Estimation of Teacher Value Added." Education Finance and Policy 4,no.4 (Fall):520-536.

Jacob, Brian A., and Lars Lefgren. 2008. "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education." Journal of Labor Economics 26,no.1 (January):101-136.

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,no.6 (December):615-631.

Kane, Thomas J., and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." Cambridge, MA, NBER w14607, National Bureau of Economic Research (December).

Koedel, Cory, and Julian R. Betts. 2009. "Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique." Columbia, MO, Department of Economics WP 09-02, University of Missouri (July).

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,no.4 (Fall):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,no.2 (May):251-266.

Nye, Barbara, Spyros Konstantopoulos, and Larry V. Hedges. 2004. "How Large Are Teacher Effects?" Educational Evaluation and Policy Analysis 26,no.3 (January):237-257.

Podgursky, Michael J., and Matthew G. Springer. 2007. "Teacher performance pay: A review." Journal of Policy Analysis and Management 26,no.4:909-949.

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,no.2 (March):417-458.

Rockoff, Jonah E. 2004. "The impact of individual teachers on student achievement: Evidence from panel data." American Economic Review 94,no.2 (May):247-252.

Rothstein, Jesse. 2009. "Student sorting and bias in value-added estimation: Selection on observables and unobservables." Education Finance and Policy 4,no.4 (Fall):537-571.

———. 2010. "Teacher quality in educational production: Tracking, decay, and student achievement." Quarterly Journal of Economics 25,no.1.

Table 1. The Distribution of Teacher Effectiveness (standard deviations of student achievement)

Study	Location	teacher effectiveness (s.d.)	
		reading	math
Rockoff (2004)	New Jersey	0.10	0.11
Nye, Konstantopoulos, and Hedges (2004)	Tennessee	0.07	0.13
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 midwest city	0.12	0.26
Kane and Staiger (2008)	Los Angeles	0.18	0.22
Koedel and Betts (2009)	San Diego		0.23
Rothstein (2010)	North Carolina	0.11	0.15
Hanushek and Rivkin (2010)	Undisclosed Texas city		0.11

Note: All estimates indicate the standard deviation of teacher effectiveness in terms of student achievement standardized to mean zero and variance one. All are corrected for test measurement error. All except Kane and Staiger (2008) use within-school estimators.