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 1
NOVEMBER 2007

Value-Added Analysis and Education Policy

STEVEN G. RIVKIN

Education researchers have recently recognized that easily quantifiable characteristics—including postgraduate education, experience, college quality, certification, and even scores on standardized tests—do not capture much of the variation in the quality of instruction as measured by the contribution to learning. This finding provides empirical support for outcome-based approaches that measure and, in some cases, reward teacher quality based on student outcomes. Such approaches eliminate the need to identify the precise relationship between teacher characteristics and the type of training undertaken by focusing on classroom performance. The No Child Left Behind (NCLB) mandate that students must be tested annually provides the necessary achievement data, and increasing numbers of districts and states are adopting pay-for-performance plans and accountability systems in which student test outcomes occupy a central role.

Despite potential shortcomings, value-added analysis can provide valuable information for evaluating and compensating teachers.

The success of outcome-based policies hinges on several factors, but the validity of the teacher quality measures is perhaps the most important. As the research literature discusses in great detail, the determinants of both student and teacher choices and the allocation of students among classrooms complicate efforts to estimate the contributions of teachers to learning. The imprecision of tests as measures of achievement, failure of some examinations to measure differences throughout the skill distribution, and limited focus of the tests on a small number of subjects further complicate efforts to rank teachers and schools based on the quality of instruction.

Yet despite these potential drawbacks, value-added analysis may still provide valuable information to use in personnel decisions and teacher compensation structures. The fact that value-added estimates will never measure precisely the quality of instruction in a classroom does not imply that they have no productive uses. Rather, recognizing the methodological issues can facilitate more informed uses of standardized test results and the development of stronger assessments.

This brief describes estimation and measurement issues relevant to estimating the quality of instruction in the context of a cumulative model of learning. It also discusses implications for the use of value-added estimates in personnel and compensation matters. The discussion highlights the importance of accounting for student differences and the advantages of focusing on student achievement gains as opposed to differences in test scores. It also recognizes, however, that the value-added framework does not address all potential impediments to consistently estimating the quality of instruction. Specific methods mitigate some problems and not others; none may resolve all potential problems. Acquiring a clearer understanding of these issues can improve the methods used to estimate added teacher value and how these estimates are used.

The next section describes student, teacher, and principal choices that are primary determinants in student-teacher matching. The following section outlines a cumulative model of learning and considers the advantages of various estimation methods given the student-teacher matching process and test measurement error and structure. The final section focuses on implications for policy. It begins by describing the advantages and disadvantages of specific methods before turning to a more general discussion of the use of value-added estimates in the teacher evaluation and compensation.

ALLOCATION OF STUDENTS AND TEACHERS INTO SCHOOLS AND CLASSROOMS

This section outlines the decision-making processes of families, teachers, and principals and the potential implications of these choices for estimating teacher quality. Families choose a community and school, possibly trading off school quality with other housing amenities. Given preferences, additional income is used in part to send children to a more expensive school (in terms of housing price capitalization, rent premium, or higher tuition in the case of a private school). Among families with the same income, higher preferences for schooling would be expected to lead families to forgo other housing amenities in order to procure a higher school quality. Because many data sets including administrative data have limited information on family income, it is often difficult to control for income differences. In addition, such data are unlikely to have information on family commitment to schooling. Because each factor likely exerts a direct effect on achievement, the failure to account adequately for family differences contaminates estimates of teacher quality.

Evidence on teacher preferences suggests that teachers tend to prefer schools with higher achieving students and have heterogeneous preferences regarding school location and characteristics related to student race and ethnicity. Survey evidence suggests that principal behavior influences the probability a teacher remains in a school, likely introducing a link between teacher and principal quality. This link likely includes the process through which principals determine classroom assignments.

A principal's objective function almost certainly influences classroom assignments and the distribution of classroom average test scores within a school. An egalitarian principal might place more disruptive children with a higher-quality teacher, while a principal that desires to please the senior staff might give experienced teachers the most compliant and skilled children. These two allocation mechanisms have very different implications for the observed achievement differences among classrooms and the estimation of teacher quality.

The purposefulness of these choices almost certainly introduces correlations among teacher quality, school quality, and family and student characteristics, thus complicating efforts to identify teacher effects on achievement. The following section develops an empirical model of achievement based on the notion that learning is cumulative. The discussion highlights how well specific methods address complications introduced by student, teacher, and principal choices as well as by test measurement error.

EMPIRICAL MODEL

Equation (1) models achievement of student i in grade G in classroom c in school s in year y as a function of student skill α_{iGy}, family background represented by X in year y, peer composition in classroom c during year y (P), school factors specific to grade G in year y including resources, principal quality, and school- or district-determined curriculum (S), teacher quality (T), and a random error. Without loss of generality, think of each term as a scalar index of the respective characteristic that increases in value as the characteristic becomes more conducive to achievement. For example, a higher value of P indicates a better peer composition (perhaps fewer disruptive students). Therefore, all the parameters are nonnegative, as higher skill, family characteristics that support achievement, better peer composition and schools, and higher-teacher quality all raise test scores.

(1)
$$A_{iGcsy} = \alpha_{iGy} + \beta X_{iGcsy} + \delta P_{Gcsy} + \delta S_{Gsy} + \lambda T_{Gcsy} + e_{iGcsy}$$

If teacher quality T were unrelated to student skill, peer quality, school quality, or other factors captured in the error, one could simply use the variation in classroom average achievement to rank teachers. In other words, if students and teachers were randomly assigned to communities, schools, and classrooms, achievement differences among classrooms would provide an unbiased ranking of teachers based on quality. Importantly, this does not mean that the ranking would constitute a true ordering by quality, because random differences in student composition and school characteristics would

contribute to classroom average achievement differences, as discussed below.

Of course, teachers and students are not randomly matched, and the various types of purposeful sorting invalidate classroom average achievement as an index of teacher quality. Consequently, such methods as regression, which isolates the effects of teacher quality from other influences, must be implemented, and the desirability of any particular approach depends on how well it accounts for the potential confounding factors.

It is unlikely that available variables account for all school and peer factors systematically related to both achievement and teacher quality.

Consider first the treatment of unobserved skill α_{iGy}. In contrast to much work that focuses on fixed ability differences emanating from early childhood experiences or other time invariant factors, the grade and year subscripts recognize explicitly that student skills evolve over time. Equation (2) captures the sense in which the full history of family, teacher, peer, community, and student influences combine to determine the student input to learning at each point in time.

(2)

$$\alpha_{iGy} = \beta \sum_{g=1}^{G-1} \theta^{G-g} X_{igy} + \delta \sum_{g=1}^{G-1} \theta^{G-g} P_{igy} + \lambda \sum_{g=1}^{G-1} \theta^{G-g} T_{igy} + (\gamma + \sum_{g=1}^{G-1} \theta^{G-g} \gamma_i)$$

A good teacher likely raises achievement in the current year and subsequent years by increasing the stock of knowledge, and a supportive parent does the same. Notice that factor effects (and knowledge) are assumed to depreciate geometrically, meaning a teacher or peer's effect on test scores diminishes with time; a good 4th grade teacher has a larger effect on a 4th grade score than on a 5th grade score. The equation does not specify the rate of deprecation.[1] If $\theta = 1$, the effects of prior experiences persist fully into the future, while if $\theta = 0$, prior experiences have no effect on current achievement. It is highly likely that the actual rate at which knowledge depreciates lies between 0 and 1.

A value-added regression of achievement in grade *G* on achievement in grade *G* - 1, family, school, and peer characteristics, and a full set of indicators for each teacher provide a natural way to account for prior influences and estimate teacher effects on achievement (the dummy variable coefficients).[2] Rewriting equations (1) and (2) for grade *G* - 1 and year *y* - 1 illustrates how the inclusion of $A_{iG-1,csy-1}$ as an explanatory variable with parameter θ in a regression with achievement in grade *G* as dependent variable potentially controls all historical factors. In the absence of test measurement error, only the contemporaneous ability effect γ_i remains unaccounted for regardless of the rate knowledge depreciates (assuming the source of knowledge does not affect the rate of depreciation). And since this fixed ability component is likely highly correlated with lagged achievement, controlling for lagged achievement differences almost certainly removes much of the variation in contemporaneous ability as well. Consequently, in the absence of measurement error, the value-added specification appears to provide an excellent method for capturing skill differences that contribute to variation in achievement among classrooms.

Although lagged achievement captures important differences among students, variation in peer composition, class size, and other school characteristics remain and are likely to be systematically related to teacher quality. This discrepancy illustrates the value of using a multiple regression framework that uses information on family characteristics, class size and other school variables, and peer variables—including average lagged test score, racial composition, and turnover (often aggregates of student characteristics)—to control for much of the remaining variation.

Nonetheless, it is unlikely that available variables account for all school and peer factors systematically related to both achievement and teacher quality, The quality of the principal, how closely the curriculum for grade *G* lines up with the state test, the level of student disruption in a school (lagged average achievement in a classroom is a crude but imperfect control for disruption), and other school and community factors are difficult to quantify, and they may contribute to systematic differences among schools in estimates of teacher effects.

In addition, the purposeful sorting of students into classrooms likely influences the distribution of achievement. A principal that wishes to equalize school quality within a grade will tend to mix more-difficult-to-educate students with better teachers, while a principal that responds to the most persistent parents or desires to please better teachers will tend to do the opposite.

Finally, consider the possibility that parents devote more time to academic support if their child has a less effective teacher. Assuming time devoted by parents for academic support is "effective," this time will tend to attenuate estimated differences in teacher effectiveness by introducing a positive correlation between the error and class size or a negative correlation between the error and true teacher effectiveness (Todd and Wolpin 2003).

All in all, it is unlikely that a value-added regression will produce unbiased estimates of teacher fixed effects (i.e., on average they will not be equal to true quality). Rather, systematic differences both within and between schools are likely to contribute to the estimates. The key issue is the magnitude of the imperfections.

Some methods can account for such unobserved factors, though in some cases they alter the nature of the comparison among teachers. Consider first the inclusion of school by grade by year fixed effects, a method that essentially compares teachers only with others in the same school, grade, and year (alternatively, one can simply include school or school by year fixed effects). This method certainly accounts for differences in myriad factors that affect all classrooms in a single grade in a year, such as a change in principal or curriculum, neighborhood improvement or decline, a change in school attendance patterns, and so on. A focus solely on within-school comparisons, however, rules out comparisons of teachers in different schools, a key aspect of many policies. This method also does not mitigate complications introduced by purposefully matching students and teachers within schools.

A second alternative is the inclusion of student fixed effects, meaning estimates of teacher quality would be based on differences in achievement gains in different grades for the same student. This approach accounts for all fixed differences among students and can substantially reduce bias from unobserved heterogeneity. But the inclusion of student fixed effects does not account for systematic differences in peer composition. And since most students remain in the same school and most students who change schools move to schools with similar demographic characteristics, student fixed effects would also alter the nature of the comparison among teachers. In addition, looking only at differences within students substantially increases demands on the data, as each student record would need three or four test scores in order to contribute to the estimation. Finally, the addition of a large number of student fixed effects potentially exacerbates test measurement–induced problems, discussed below.

TEST MEASUREMENT ISSUES

A crucial consideration in empirical analysis of student achievement is that achievement as measured by a given standardized test rarely if ever equates exactly with the conceptual notion of achievement as the level of mastery in a particular academic area. First, all tests measure knowledge with error—that is, the score reflects a combination of knowledge, luck, and whether the test-taker had a good day. Measurement error in the dependent variable does not bias estimates of teacher effects, but measurement error in the lagged test score used as a control biases the coefficient on the test score variable and potentially on the other variables as well. In both cases, measurement error increases standard errors. Second, tests inevitably emphasize some skills more than others, and this emphasis can distort the results. Curricular differences among schools and districts influence the time allocated to each subject and, therefore, knowledge of particular material. Consequently, test coverage affects examination results, and even unbiased estimates of teacher value added do not necessarily index quality; it might take some teachers far more time than others to produce a given amount of learning. Such differences in time allocation would affect learning in nontested subjects.

Random error in the grade G test score leads to errors in ranking teachers and schools based on their true impact on knowledge measured by the tests. This problem can have financial implications in states that use test scores to reward schools and teachers, and it can provide potentially misleading information to administrators making employment decisions and mentoring teachers. A particular striking example provided by Kane and Staiger (2001) is the much higher probability that quality estimates for schools or teachers with small numbers of students will fall in the tails of the distribution, meaning reward or punishment systems that focus on those at the top or bottom are likely to disproportionately reward or punish low-enrollment schools or teachers. In general, it is important to recognize that such noise exists when using the tests in high-stakes situations, to learn about the reliability of tests, and to mitigate the magnitude of the error where possible.

Several mechanisms exist for reducing the influence of measurement error in the outcome variable, including the use of adaptive tests that yield more precise estimates of knowledge for a given number of questions, increases in the number of test items, and increases in the number of students tested per teacher. Increasing the number of test items has a cost in terms of student time. Raising the number of students per teacher is limited by enrollment, though a system could use multiple years of available information to estimate teacher effects.

Improving tests and adding items also makes lagged achievement a better measure of accumulated knowledge, and researchers can add other tests as controls, including those from previous years or other subjects. As Ballou, Sanders, and Wright (2004) discuss, Bayesian shrinkage estimators can be used with multiple lagged tests to produce more precise estimates of teacher effects. Differences in test availability that vary systematically among teachers can complicate the use of such methods in trying to rank teachers across schools and districts.

VALUE ADDED AND POLICY

The myriad factors that influence cognitive growth over an extended period, the purposeful sorting of families and teachers into schools and classrooms, and the imperfections of tests as measures of knowledge complicate efforts to estimate teacher fixed effects and rank teachers according to quality of instruction. Yet despite potential shortcomings, value-added analysis can provide valuable information for use in evaluating and compensating teachers. The key is not to be cavalier about the information contained in value-added estimates but to understand the pieces that go into producing estimates of teacher quality.

The myriad factors that influence cognitive growth, the purposeful sorting of families and teachers into schools and classrooms, and the imperfections of tests as measures of knowledge complicate efforts to estimate teacher effects.

A common question is whether to focus on outcomes at the teacher or school level. The best answer might be that the appropriate level depends on the issue at hand. Outcomes at the teacher level inform teachers and administrators about the quality of instruction and the success of specific pedagogical methods. Principals have first-hand knowledge about the classroom (including information not available for statistical analysis, such as time devoted to the tested material and classroom composition) that can contextualize the results. Such information can provide supportive evidence and strengthen the principal's hand in efforts to remove ineffective teachers or require teachers to undergo remediation. More generally, value-added estimates provide a benchmark against which principals can compare subjectively formed opinions of teacher effectiveness. Of course, the better a principal understands value-added analysis, the more effectively she is likely to be able to use the information to improve the quality of instruction in her school.

One important question is whether to compute value added from a single year of test results or to average over multiple years. Averaging teacher performance over a number of years increases the precision of the estimates but dampens the incentives to improve: a bad year will weigh down future assessments and lessen the reward when a teacher recovers from what might have been a difficult entry into the profession or a bad period due to personal problems or other factors. In addition, differences in the number of years of available test results complicate schemes that use average results.

In cases where principal incentives are not well aligned with the production of knowledge or bargaining constraints limit principal flexibility, value-added estimates can provide the basis for personnel decisions designed to improve the quality of instruction. For example, consider a framework that automatically fires the bottom x percent of teachers in terms of value added in a single year or average value added over a number of years. Unobserved differences among schools and classrooms almost certainly influence the estimates, and test error certainly introduces a degree of randomness. Consequently, mistakes will be made; however, outcomes could still improve compared with the system without such decision rules. Nonetheless, the implications of adding such risk and uncertainty may necessitate a substantial salary increase, and these monetary costs as well as costs associated with increased turnover would have to be weighed against any improvements in the composition of teachers.

Linking compensation to value added at the teacher level is more complicated for

several reasons. Of particular importance is the matching process used to assign students to classrooms. The use of teacher test score results to award performance pay might compromise the ability of principals to balance the educational experience for all students. It would be unfair to assign more difficult students to the more effective teachers if so doing would reduce the compensation of those teachers. In addition, it is difficult to account for differences among schools and peer groups, meaning teachers in well-run schools with peer groups that facilitate learning will have an advantage.

Average value added in a single grade appears to constitute a more fruitful approach; it provides incentives to teachers and school administrators and information to families despite difficulties in fully accounting for community and family differences that contribute to variation in school average achievement. Although cooperation among teachers is a worthy consideration, there are at least two more important reasons to focus on school-level outcomes. First, this focus does not impede principals from considering the strengths and weaknesses of teachers and students in classroom allocation. Second, this focus provides a strong set of incentives for school leaders who make the key personnel, spending, and curricular decisions and should be held accountable for their actions.

Regardless of whether the focus is on schools or individual teachers, the types of estimation and testing issues raised throughout this brief should be kept in mind. A key problem introduced by the fact that tests measure knowledge with error is that the probability of being at the top or bottom ranking varies inversely with the number of students tested. Consequently, compensation and sanctioning systems that focus on the small number of schools or teachers at the top or bottom are likely to provide very weak incentives to schools with higher enrollments and thus a much lower chance of being in one tail of the distribution.

An alternative approach that circumvents such problems is to spread the rewards over the entire range of value added, such as a system in which additional compensation is a continuous variable that varies inversely with estimated value added. This system creates incentives for all teachers and schools regardless of size. As long as the scheme is symmetric (the loss from being at the bottom as opposed to the middle equals the gain from being at the top as opposed to the middle), expected reward should not differ systematically by school size. Of course, programs that provide large rewards for small numbers of schools or teachers may produce stronger incentives, creating a potential trade-off between incentive strength and fairness.

All in all, care should be used in the determination of the appropriate role of accountability systems and standardized tests. There is a temptation to substitute a mechanical, standardized performance system in place of administrator evaluations of teachers in much the same way as standardized teacher licensing tests are used to limit or eliminate administrator discretion in teacher hiring. As Dale Ballou pointed out, licensing test cutoffs limit the use of the entire body of information provided to administrators and may in fact diminish the average quality of new hires.[3] All these tests provide abundant information for administrator use and, in the case of student accountability examinations, for the construction of school performance measures. Local administrators should be given enough flexibility and control so they can take ownership over the school and be responsible for its success or failure.

NOTES

1. An alternative value-added specification is to use the difference in scores between grades G and G - 1 as the dependent variable, thus imposing the assumption of $\theta = 1$. As Rivkin demonstrates, this more restrictive framework will tend to bias downward differences among teachers in the absence of student fixed effects and bias upward differences among teachers if student fixed effects are included (Steven Rivkin, "Cumulative Nature of Learning and Specification Bias in Education Research," unpublished manuscript, 2006).

2. See Hanushek (1986) for a discussion of value-added models.

3. Dale Ballou, personal correspondence with the author.

REFERENCES

Ballou, Dale, William Sanders, and Paul Wright. 2004. "Controlling for Student Background in Value-Added Assessment of Teachers." *Journal of Educational and Behavioral Statistics* 29(1): 37–65.

Hanushek, Eric A. 1986. "The Economics of Schooling: Production and Efficiency in Public Schools." *Journal of Economic Literature* 24(3): 1141–77.

Kane, Thomas J., and Douglas O. Staiger. 2001. "Improving School Accountability Measures." Working Paper 8156. Cambridge, MA: National Bureau of Economic Research.

Todd, Petra E, and Kenneth I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *Economic Journal* 113(485): F3–F33.

ABOUT THE AUTHOR

Steven G. Rivkin is a professor of economics at Amherst College. His research focuses on topics in the economics and sociology of education that tend to closely relate to current policy discussions. Dr. Rivkin is also associate director of research for the Texas Schools Project at the University of Texas, Dallas, a fellow at the National Bureau of Economic Research, and a CALDER researcher. His full C.V. is available at http://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

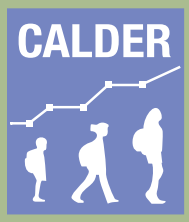
IN THIS ISSUE

Value-Added Analysis and Education Policy

For more information, call Public Affairs at 202-261-5709 or visit our web site, http://www.urban.org.

To order additional copies of this publication, call 202-261-5687 or 877-UIPRESS, or visit our online bookstore, http://www.uipress.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.

The author thanks Dale Ballou for very helpful discussions on a preliminary draft. The author also thanks Kim Rueben, Elizabeth Cohen, and the Urban Institute.