Statement on the New York City Teacher Value-Added Model
Value-Added Data for Teachers Research Team

August 29, 2008

Overview

In developing value added measures, NYC DOE relied on a number of academic experts as technical advisors. It is the view of the experts listed below that the teacher value-added measures being proposed by NYC DOE provide useful information regarding individual teacher's effectiveness in the classroom. Our support for the teacher value-added model being proposed by NYC is limited to the technical quality of the work and its promise for informing improvements in teaching and learning, while not endorsing any particular use for accountability, promotion, or tenure. The teacher value-added model being proposed by NYC uses methodology that is similar to that used in peer-reviewed research in academic journals and by other school districts to evaluate teacher effectiveness in raising student achievement. While these value-added measures have well-known limitations, and some important questions remain about how to most effectively use and interpret these measures, they provide useful objective information that is not readily available from other sources. Sharing this information with administrators and teachers will both provide school staff additional information for improving their practice and help NYC DOE to refine their methodology and to learn about how to use this type of information to raise student achievement.

Brief Description of NYC DOE Teacher Value-Added Model

In simple language, the teacher value-added model that has been developed by NYC DOE is based on the extent to which each student's test score exceeded or fell behind expectations.[1] The expected score for each student is based on the average score that was observed among students with similar characteristics at the beginning of the year (e.g., students who had similar test scores in the prior year and who were similar in terms of demographics and program participation) and who were in classrooms and schools with similar student characteristics. Each teacher's value added estimate is the average difference between the actual test score and the expected score across all of the students assigned to that teacher. This is done using only the most recent year of data and using multiple years, and is done separately by subject (Mathematics and English Language Arts) and by grade (4-8). Finally, value added at the teacher level is compared to two groups of teachers: all teachers in the same grade, and teachers in the same grade with similar experience and students with similar characteristics.

Because value added estimates for any single teacher are based on a limited number of years and classrooms, they will overstate or understate the effectiveness of some teachers because of chance events. While some error is inevitable, the NYC DOE methodology incorporates a commonly used statistical adjustment (referred to as Empirical Bayes or shrinkage techniques) that mitigates the impact of chance on value added estimates. This adjustment is incorporated into the calculation of the expected gain for each teacher.

This general approach to estimating value added has been used widely in prior research to compare the effectiveness of teachers with different levels of experience,

education and other credentials[2] and to estimate the differences in effectiveness across individual teachers.[3] This approach has also been used by other school districts and other countries to evaluate the performance of teachers and schools.[4] Similar methods have been used to evaluate the performance of hospitals and physicians in the U.S. and elsewhere, for both research and public reporting purposes.[5] There is a growing literature documenting the statistical properties of value-added measures, which suggests that the approach being taken by NYC DOE is a reasonable approach that will yield estimates of teacher value added that are similar to most other common approaches.[6] The NYC DOE methodology is similar in spirit to the well-known value-added methodology used by Tennessee to evaluate teacher effectiveness, but differs substantially in the details. Two key differences are that the NYC methodology (1) adjusts for more factors in computing the expected gain for each student and (2) only uses scores from tests taken while the student was a member of a teacher's class, rather than information from the student's testing trajectory over their entire school experience.

Because tests in New York are currently administered prior to the end of the school year (January for English language Arts, and March for Mathematics), there are important concerns that test scores in one year may reflect the quality of instruction a student received during the latter part of the prior year. To our knowledge, there is no extant research focusing on this topic and how to effectively deal with it in a value-added framework (other studies have either used tests from late in the spring or ignored the issue of mid-year testing). The methodology proposed by NYC DOE is a reasonable first step to addressing the issue of mid-year testing. To estimate the value added of each teacher, the expected gain of each student is adjusted to account for the teacher that he or she had in the prior school year. This adjustment is, of course, in addition to the previously mentioned adjustments for student, classroom and school characteristics. Thus, a teacher with high value added in the testing year is one whose students had higher test gains than did students who, in addition to having similar characteristics, had the same teacher in the prior year.

Strengths of the Value-Added Measures

The value-added measures that have been developed by NYC DOE have a number of important advantages. First, the methodology is relatively simple and transparent. This allows users (teachers and administrators) to better evaluate and provide feedback on the validity of the data they receive. For example, users have access to the adjustments that generated the expected gains for each student, and can consider whether these adjustments generated plausible differences in expected gains between their students or if there was some important factor that the model had not incorporated. This simple methodology also allows users of these measures to "drill down" and compare actual to expected performance for subsets of students within a classroom. This will encourage exploration into reasons for strong or weak performance among particular groups or students.

A second strength of these measures is that they are based on a value-added methodology that has been validated in a number of important ways. Other measures of teacher performance have been found to be positively related to teacher value added, ranging from teacher experience and certification test scores to those derived from classroom observation, principal evaluations, and parent feedback.[7] At the same time,

value-added measures capture additional information that is not in these other measures – for example, compared to other measures of teacher performance, current estimates of teacher value added are better able to predict the future impact of a teacher on student's test scores.[8] Experimental studies, in which teachers were randomly assigned to classrooms, have validated the value-added approach as well. Experimental estimates of the impact of Teach for America and National Board Certified teachers on student achievement were similar to observational estimates based on the value-added approach.[9] In a recent study of elementary teachers in Los Angeles, estimates of teacher value-added (using an approach very similar to that used by NYC DOE) based on pre-experimental data were found to be accurate predictors of student achievement differences across teachers following random assignment – when a teacher with higher value added was randomly assigned to a class, the students in that class had higher test scores at the end of the year.[10] While this study must be replicated elsewhere before reaching any definitive conclusion, it suggests that estimates of teacher value added such as those being proposed by NYC DOE provide useful information.

Limitations of the Value-Added Measures and Questions to Address Going Forward

As with any single measure of teacher effectiveness, there are a number of important limitations of current value-added measures that must be kept in mind. Some of these limitations will require ongoing investigation:[11]

1) Test scores capture only one dimension of teacher effectiveness, and they are not intended to serve as a summary measure of teacher performance. Obviously, they do not measure a student's academic achievement in untested subjects or grades. In addition, they do not measure valued non-cognitive skills (e.g., classroom behavior, study skills). Therefore, value-added measures should only be considered as one dimension of teacher effectiveness.
2) If high stakes are attached, there will be potential to game these measures by teaching to the test, selecting students, altering difficult-to-audit student characteristics, or outright cheating. Of course, some distortions can be expected with any high stakes measure. These distortions must be monitored, and will be lessened if value-added is one of many measures used to evaluate teacher effectiveness.
3) To calculate expected test scores, NYC DOE considered a wide range of factors and included those that were significant predictors of student gains. Nevertheless, there are likely to be additional factors not yet considered that influence student achievement. If high stakes are attached to the value added measures, there will be an increased incentive to game the measures through purposeful assignment of students based on such omitted factors. Therefore, to the extent feasible, additional factors identified by researchers, teachers, or school administrators should be considered and added to the model if they have a substantive impact on estimates of teacher value added.
4) As with all measures, value added estimates are imperfect. They may overstate or understate the effectiveness of individual teachers because of chance events or because the teacher's effectiveness has changed since the time they were last observed. NYC DOE plans to provide information on this uncertainty through reporting of confidence bands for the value added estimates. It is important that users understand that all measures of teacher performance (value added included) are

imperfect, and are best interpreted in the context of other information available to the user.

5) Because of chance events, value added estimates based on data from one or two classrooms will be unstable over time even if true teacher effectiveness remains unchanged.[12] It is important that users be made aware of likely instability particularly when reporting estimates based on a single class. Developing methods that more systematically incorporate past information on value added into current estimates will help to eliminate this instability and should be a priority in the future.
6) The method of adjusting expected student gains to account for prior teacher (because of the mid-year testing dates) is new. Validating this methodology, and considering other possible alternatives, should be a high priority.
7) Most of the validation evidence for value added measures come from within-school comparisons of teachers. The evidence base supporting value added for comparing teachers across schools, especially across schools serving very different populations, is less strong. In particular, all of the experimental studies randomized teachers within a school, not between schools. Future work should establish that teachers with high value added in one school would continue to be high value added if they taught in a different school.
8) A growing body of evidence from experimental and non-experimental studies has found that the difference between being assigned an effective and ineffective teacher is largest in the short term (e.g., on end-of-year test scores) but tends to be more muted in the longer term (e.g., on test scores 2 years later).[13] However, it is not clear what should be made of such "fade out" effects. Obviously, it would be troubling if students are simply forgetting what they have learned, or if value-added measures something transitory (like teaching to the test) rather than true learning. This would imply that value added overstates teacher effectiveness. However, this "fade out" evidence could also reflect changing content of the tests in later grades, or that the impact of a good teacher spills over to other students in future years through peer effects. Better understanding of the nature of the fade-out is needed before concluding that teacher effects on student achievement are ephemeral.

Conclusions

The teacher value-added model being proposed by NYC DOE uses methodology that is similar to that used in academic research and by other school districts. These value-added measures have well-known limitations and some important questions remain about how to most effectively use and interpret these measures. However, they also provide useful objective information that is not readily available from other sources.

The DOE is considering providing the information to all eligible schools. We think this is a worthwhile activity, provided that administrators and teachers receive detailed training on the limitations of the data and about sound approaches to its interpretation. Sharing this information with administrators and teachers will both provide school staff additional information for improving their practice and help NYC DOE to refine their methodology and to learn about how to best use this type of information to raise student achievement.

Finally, the teacher value-added model being proposed by NYC DOE is clearly a first step. We encourage the DOE to develop specific plans to further refine and validate these measures, and to train educators in their use. Such a process would establish an evidence base that was more specific to NYC schools, would provide skeptics with more confidence in the measures, and would potentially establish the NYC DOE value added system as a model for the country.

Technical Advisors

Thomas J. Kane, Harvard University
Jon Fullerton, Harvard University
Jonah Rockoff, Columbia University
Douglas Staiger, Dartmouth College

---

[1] For details on the methodology, see the technical report from NYC DOE.
[2] For recent examples, see Boyd et al. (2006); Kane, Rockoff and Staiger (2007); Clotfelter, Ladd and Vigdor (2007), Harris and Sass (2006).
[3] For recent examples, see Rockoff (2004); Hanushek, Kain, O'Brian, and Rivkin (2005); Rivkin, Hanushek and Kain (2005); Murnane, Willett, Somers and Uribe (2005), Aaronson, Barrow and Sanders (2007); Gordon, Kane and Staiger (2006), Koedel (2007). See also Hoffman and Oreopoulos (2006) and Carrell and West (2008) on this issue at the tertiary level.
[4] Some form of value added is currently being used in districts in North Carolina, Ohio, Pennsylvania, Tennessee, and Texas (see McCaffrey and Hamilton, 2007 for a recent survey). For use of value added in England, see Evans (2008).
[5] For recent overviews, see Birkmeyer and Birkmeyer (2006), Marshal et al. (2003) and Normand and Shahain (2007).
[6] See McCaffrey et al. (2003) and the articles in Journal of Educational and Behavioral Statistics, Vol. 29, no. 1-2 (Value Added Assessment Special Issue).
[7] For recent examples, see Daley and Valdes (2006), Gallagher (2004), Jacob and Lefgren (2005), Kimball et al. (2004), Milanowski (2004), and Harris and Sass (2007).
[8] Cantrell et al. (2007), Jacob and Lefgren (2005).
[9] See Cantrell et al. (2007) and Decker, Mayer and Glazerman (2004).
[10] Kane and Staiger, 2008.
[11] See Richard Rothstein (2008) for an excellent summary of the limitations of value added.
[12] See Kane and Staiger (2002) and McCaffrey, Sass and Lockwood (2008).
[13] For U.S. examples, see Jacob et al. (2008), Kane and Staiger (2008), Konstantopoulos (2007,2008), Krueger and Whitmore (2001), McCaffrey et al. (2003) and Jesse Rothstein (2008). In developing countries see Banerjee et al. (2007) and Glewwe et al. (2003).

References

Aaronson, D., L. Barrow, and W. Sander, "Teachers and Student Achievement in the Chicago Public High Schools," Journal of Labor Economics, 25:1, 95-135, 2007.

Banerjee, A.V., S. Cole, E. Duflo and L. Linden, "Remedying Education: Evidence from Two Randomized Experiments in India," Quarterly Journal of Economics, August 2007, Vol. 122, No. 3, Pages 1235-1264.

Birkmeyer, N.J.O. and J.D. Birkmeyer, "Strategies for Improving Surgical Quality – Should Payers Reward Excellence or Effort?" New England Journal of Medicine, 354(8):864-870, Feb. 23 2006.

Boyd, D., P. Grossman, H. Lankford, S. Loeb, and J. Wyckoff, "How changes in Entry Requirements Alter the Teacher Workforce and Affect Student Achievement," Education Finance and Policy, Spring 2006, Vol. 1, No. 2, pages 176-216.

Cantrell, S., J. Fullerton, T.J. Kane, and D.O. Staiger, "National Board Certification and Teacher Effectiveness: Evidence from a Random Assignment Experiment," working paper, March 2007.

Carrell, S.E. and J. West, "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors," NBER Working Paper 14081, June 2008.

Clotfelter, C.T., H.F. Ladd and J.L. Vigdor, "How and Why do Teacher Credentials Matter for Student Achievement," NBER working paper #12828, January 2007.

Daley, Glenn and Rosa Valdés. 2006. Value Added Analysis and Classroom Observation as Measures of Teacher Performance: A Preliminary Report. Los Angeles Unified School District.

Decker, Paul T., Daniel P. Mayer and Steven Glazerman. (2004) "The Effects of Teach For America on Students: Findings from a National Evaluation," Mathematica Policy Research Report No. 8792-750, June 9, 2004.

Evans, H., "Value Added in English Schools," manuscript prepared for the National Conference on Value-Added Modeling, Madison, WI, April 2008 (paper available at http://www.wcer.wisc.edu/news/events/VAM%20Conference%20Final%20Papers/VAM inEnglishSchools_HEvens.pdf)

Gallagher, H. Alix. 2004. "Vaughn Elementary's Innovative Teacher Evaluation System: Are Teacher Evaluation Scores Related to Growth in Student Achievement?" *Peabody Journal of Education* 79(4): 79-107.

Glewwe, P., N. Ilias, and M. Kremer, "Teacher Incentives," NBER working paper #9671, May 2003.

Gordon, R., T.J. Kane, and D.O. Staiger, "Identifying effective teachers using performance on the job," Hamilton Project discussion paper, Washington DC: Brookings Institution.

Hanushek, E.A., J.F. Kain, D. O'Brien, and S.G. Rivkin, "The Market for Teacher Quality," NBER working paper #11154, 2005.

Harris, D. and T.R. Sass, "The Effects of Teacher Training on Teacher Value Added" Florida State University Working Paper, 2006.

Harris, D. and T.R. Sass, "What Makes for a Good Teacher and Who Can Tell?" Florida State University Working Paper, 2007.

Hoffman, F. and P. Oreopoulos, "Professor Qualities and Student Achievement," NBER Working Paper 12596, October 2006.

Jacob, Brian A. and Lars Lefgren. 2005. "Principals as Agents: Subjective Performance Measurement in Education." NBER working paper #11463, July 2005.

Jacob, B.A., L. Lefgren, and D. Sims, "The Persistence of Teacher-Induced Learning Gains," NBER working paper #14065, June 2008.

JEBS special issue, vol 29, no 1-2, value added assessment special issue.
Journal of Educational and Behavioral Statistics, Vol. 29, no. 1-2, Value Added Assessment Special Issue, 2004.

Kane, T.J. and D.O. Staiger, "The Promise and Pitfalls of Using Imprecise School Accountability Measures," Journal of Economic Perspectives, 16(4):91-114, Fall 2002.

Kane, T.J., J. E. Rockoff and D.O. Staiger, "What does certification tell us about teacher effectiveness? Evidence from New York City," Economics of Education Review (2007), doi:10.1016/j.econedurev.2007.05.005.

Kane, T.J. and D.O. Staiger, "Are Teacher Value-Added Estimates Biased? An Experimental Validation of Non-Experimental Teacher Effects," working paper, March 2008.

Kimball, Steven M., Brad White, and Anthony T. Milanowski. 2004. "Examining the Relationship between Teacher Evaluation and Student Assessment Results in Washoe County." *Peabody Journal of* Education 79(4): 54-78.

Koedel, C. "Teacher Quality and Educational Production in Secondary School," University of Missouri Working Paper, 2007.

Konstantopoulos, S. "How Long Do Teacher Effects Persist?" IZA Discussion Paper No. 2893, June 2007.

Konstantopoulos, Spyros (2008) "Do Small Classes Reduce the Achievement Gap between Low and High Achievers? Evidence from Project STAR" *The Elementary School Journal* Vol 108, No. 4, pp. 278-291.

Krueger, A.B. and D.M. Whitmore, "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR," Economic Journal, 111(January):1-28, 2001.

Marshal, M.N., P.G. Shekelle, H.T.O. Davies, and P.C. Smith, "Public Reporting on Quality in the United States and the United Kingdom," Health Affairs, 22(3):134-148, May/June 2003.

McCaffrey, Daniel, J.R. Lockwood, Daniel Koretz and Laura Hamilton (2003) Evaluating Value-Added Models for Teacher Accountability, (Santa Monica, CA: Rand Corporation).

McCaffrey, D.F. and L.S. Hamilton, "Value-Added Assessment in Practice," RAND Technical Report, The RAND Corporation, Santa Monica, CA, 2007.

McCaffrey, D.F., T.R. Sass and J.R. Lockwood, "The Intertemporal Stability of Teacher Effect Estimates," Working paper, June 27, 2008.

Milanowski, Anthony. 2004. "The Relationship between Teacher Performance Evaluation Scores and Student Achievement: Evidence from Cincinnati." *Peabody Journal of Education* 79(4): 33-53.

Murnane, R.J., J.B. Willett, M. Somers, and C. Uribe "Learning Why More Learning Takes Place in Some Classrooms Than Others," German Economic Review 6(3): 309–330, August 2005.

Normand, S.L.T. and D.M. Shahain, "Statistical and Clinical Aspects of Hospital Outcomes Profiling," Statistical Science, 22(2):206-226, 2007.

Rivkin, S.G., E.A. Hanushek, and J.F. Kain, "Teachers, Schools, and Academic Achievement," Econometrica 73 (2), 417–458, 2005.

Rockoff, J.E., "The impact of individual teachers on student achievement: Evidence from panel data," American Economic Review, 94(2), 247-252.

Rothstein, J., "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," Princeton University Working Paper, May 2008.

Rothstein, R., "Holding Accountability to Account: How Scholarship and Experience in Other Fields Inform Exploration of Performance Incentives in Education," working paper 2008-04, National Center on Performance Incentives, Nashville, TN, 2008.