EDUCATION POLICY FOR ACTION SERIES

EDUCATION CHALLENGES FACING NEW YORK CITY

EXECUTIVE SUMMARY

Can Teachers be Evaluated by their Students' Test Scores? Should They Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice

Sean P. Corcoran

in collaboration with
Annenberg Institute research staff

Annenberg Institute for School Reform

AT BROWN UNIVERSITY

EDUCATION POLICY FOR ACTION SERIES
EDUCATION CHALLENGES FACING NEW YORK CITY

EXECUTIVE SUMMARY

Can Teachers be Evaluated by their Students' Test Scores? Should They Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice

Sean P. Corcoran

in collaboration with
Annenberg Institute research staff

About the Annenberg Institute for School Reform

The Annenberg Institute for School Reform is a national policy-research and reform-support organization, affiliated with Brown University, that focuses on improving conditions and outcomes for all students in urban public schools, especially those serving disadvantaged children. The Institute's vision is the transformation of traditional school systems into "smart education systems" that develop and integrate high-quality learning opportunities in all areas of students' lives – at school, at home, and in the community.

The Institute conducts research; works with a variety of partners committed to educational improvement to build capacity in school districts and communities; and shares its work through print and Web publications. Rather than providing a specific reform design or model to be implemented, the Institute's approach is to offer an array of tools and strategies to help districts and communities strengthen their local capacity to provide and sustain high-quality education for all students.

A goal of the Institute is to stimulate debate in the field on matters of important consequence for national education policy. This report provides one such perspective but it does not necessarily reflect the opinion of the Annenberg Institute for School Reform.

Annenberg Institute for School Reform at Brown University

Box 1985
Providence, Rhode Island 02912

233 Broadway, Suite 720
New York, New York 10279

www.annenberginstitute.org

About the Author

Sean P. Corcoran is an assistant professor of educational economics at New York University's Steinhardt School of Culture, Education, and Human Development, an affiliated faculty of the Robert F. Wagner Graduate School of Public Service, and a research fellow at the Institute for Education and Social Policy (IESP). He has been a research associate of the Economic Policy Institute in Washington, D.C., since 2004 and was selected as a resident visiting scholar at the Russell Sage Foundation in 2005-2006. In addition to being a member of the board of directors of the Association for Education Finance and Policy (formerly the American Education Finance Association), he serves on the editorial board of the journal *Education Finance and Policy*.

Corcoran's research focuses on three areas: human capital in the teaching profession, education finance, and school choice. His recent papers have examined long-run trends in the quality of teachers, the impact of income inequality and court-ordered school finance reform on the level and equity of education funding in the United States, and the political economy of school choice reforms. In 2009, he led the first evaluation of the Aspiring Principals Program in New York City, and he is currently working on a retrospective assessment of the Bloomberg-Klein reforms to school choice and competition in New York City for the American Institutes for Research. He co-edits a book series on alternative teacher compensation systems for the Economic Policy Institute, and in recent years he has been interested in value-added measures of evaluating teacher effectiveness, both their statistical properties and their obstacles to practical implementation.

His recent publications can be found in the *Journal of Policy Analysis and Management,* the *Journal of Urban Economics, Education Finance and Policy,* and the *American Economic Review*.

About the Series

Education Policy for Action: Education Challenges Facing New York City is a series of research and policy analyses by scholars in fields such as education, economics, public policy, and child welfare in collaboration with staff from the Annenberg Institute for School Reform and members of a broadly defined education community. Papers in this series are the product of research based on the Institute's large library of local and national public education databases; work with the Institute's data analysis team; and questions raised and conclusions drawn during a public presentation and conversation with university and public school students, teachers, foundation representatives, policy advocates, education reporters, news analysts, parents, youth, and community leaders.

Among the issues that the series addresses are several pressing topics that have emerged from the Institute's research and organizing efforts. Some of the topics covered in the series are:

- Confronting the impending graduation crisis
- The small schools experiment in New York City
- Positive behavior and student social and emotional support
- Modes of new teacher and principal induction and evaluation

Many thanks to the Robert Sterling Clark Foundation for its support of the public conversations from which this report and the other publications in the series grew.

For a downloadable version of this report and more information about the series, please visit <www.annenberginstitute.org/WeDo/NYC_Conversations.php>.

Acknowledgments

I thank the Annenberg Institute for School Reform for the invitation to conduct this research and write this report. Deinya Phenix was an immense help from start to finish. Norm Fruchter, Ivonne Garcia, Megan Hester, Christina Mokhtar, and Eric Zachary offered thoughtful and constructive feedback at multiple points during the process of writing and preparing for my January 27, 2010, presentation, which was part of the Education Policy for Action conversation series. Many audience members at this event offered insightful thoughts and comments, and I would particularly like to express my appreciation to Leo Casey from the United Federation of Teachers for serving as a discussant.

I would also like to thank Jennifer Jennings and Andrew Beveridge for sharing data from the Houston Independent School District, Rhonda Rosenberg and Jackie Bennett of the United Federation of Teachers for providing assistance with the New York City Teacher Data Report data, and Rachel Cole for her research assistance. Amy McIntosh and Joanna Cannon of the New York City Department of Education were instrumental in providing background on the Teacher Data Initiative and assisted me in correcting factual errors. All remaining errors are my own.

This Executive Summary briefly describes the findings presented in the full report *Can Teachers be Evaluated by their Students' Test Scores? Should They Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice*. Data from the study and more detailed findings are provided in the full report. For more information about the study and to download the full report or the Executive Summary, please visit <www.annenberginstitute.org/Products/Corcoran.php>.

Introduction

"Value-added" measures of teacher effectiveness are the centerpiece of a national movement to evaluate, promote, compensate, and dismiss teachers based in part on their students' test results. Federal, state, and local policy-makers have embraced these measures in recent years as a means to objectively quantify teacher quality and to identify, reward, and retain teachers with a demonstrated record of success. For example, in New York City, the Department of Education now releases "Teacher Data Reports" to its teachers in grades four to eight that concisely summarize teachers' value-added information. In Washington, D.C., and Houston, teachers can be granted or denied tenure partially based on value-added, and Houston awards bonuses to its high value-added teachers.

In theory, a teacher's value-added is the unique contribution she makes to her students' academic progress – that is, the portion of her students' achievement that cannot be attributed to any other current or past student, family, teacher, school, peer, or community influence. Because students are rarely assigned randomly to teachers, value-added measures must rely on complex statistical models to infer how much better or worse a student performed under one teacher than they would have performed under another. The ultimate goal of these tools, then, is to differentiate the causal impact of individual teachers on student outcomes.

Few can deny the intuitive appeal of value-added assessment: if a statistical model can isolate a teacher's unique effect on achievement, the possibilities seem endless. Teacher quality is an immensely important resource, and research has found that teachers can and do vary widely in their effectiveness (e.g., Kane, Rockoff & Staiger 2008). Common measures of teacher qualifications, such as experience and college selectivity, typically provide minimal information about individual teachers' effectiveness. Value-added holds out the promise that the elusive concept of "teacher quality" can be objectively and precisely measured.

However, these tools have limitations and shortcomings that are not always apparent to interested stakeholders – including teachers, principals, and policy-makers – or even to value-added advocates. In the report this executive summary is based on, I provide an introduction to these new measures of teaching effectiveness; describe prominent value-added systems currently in use in New York City and Houston; assess the potential for value-added measurement to improve student outcomes, using these programs as empirical case studies; and outline some important challenges facing their implementation in practice. This executive summary summarizes these concepts and findings; for more detailed background on the New York City and Houston programs, data analysis, and discussion, see the full report at <www.annenberginstitute.org/Products/Corcoran.php>.

What is a Teacher's Value-Added?

A teacher's value-added can be thought of as her students' average test scores "properly adjusted" for the effects of other influences on achievement. For example, in New York City's Teacher Data Reports, students' actual scores under a given teacher are compared to their predicted score – that is, their predicted achievement had they been taught by another teacher in the district (say, the average teacher). This prediction is based on a number of factors, the most important of which is the student's prior achievement. Because the school district has richly detailed data on thousands of students' academic histories, it can provide a statistical estimate of how each student is likely to have performed on a test given their background characteristics. How a student actually performs under a teacher relative to this prediction is the teacher's value-added for that student.

Though not always obvious to most observers, value-added in practice is a relative concept. It tells us how teachers measure up when compared with other teachers in the district or state with similar students. On the New York City Teacher Data Report, this is reported as the teacher's percentile in the distribution of teachers with similar experience in the same grade and subject. For example, based on last year's test results, an eighth-grade math teacher's value-added might place him at the 43rd percentile citywide. In other words, 43 percent of teachers had lower value-added than he did (and 57 percent had higher value-added). This percentile is then mapped to one of five performance categories ("high," "above average," "average," "below average," and "low"). New York City has recently encouraged principals to use these metrics in making teacher tenure decisions. In Houston's ASPIRE (Accelerating Student Progress, Increasing Results & Expectations) program, value-added measures are used in tenure decisions as well as in a system of bonus payments; teachers scoring in the top performance categories are awarded bonuses as high as $10,300 per year.

Because value-added is statistically estimated, it is subject to uncertainty, or a "margin of error." On the Teacher Data Report, this is reported as a range of possible percentiles associated with the value-added score (also called the percentile's "confidence interval"). For example, a teacher at the 43rd percentile might have a range that extends from the 15th percentile to the 71st. This means that the statistical model cannot rule out the possibility that this teacher falls somewhere between the 15th and 71st percentiles, although the 43rd is her "most likely" ranking. New York's reports include value-added measures, percentiles, and performance categories for all tested students, as well as for subgroups of students, such as initially low-achieving students or English language learners (ELLs).

Challenges to the Practical Implementation of Value-Added

I categorize the conceptual and practical challenges to value-added methods of evaluating teachers into six key questions:

- What is being measured?
- Is the measurement tool appropriate?
- Can a teacher's unique effect be isolated?
- Who counts?
- Are value-added scores precise enough to be useful?
- Is value-added stable from year to year?

What is being measured?

Value-added measurement works best when students receive a single numeric test score every year on a continuous developmental scale – that is, one that does not depend on grade-specific content, but rather progresses across multiple grade levels. The set of skills and subjects that can be adequately assessed in this way is remarkably small. Not all subjects are or can be tested, and even within tested subject areas, only certain skills readily conform to standardized testing. Yet, valid value-added measures depend entirely on such tests. Houston's ASPIRE program currently incorporates results from two sets of core subject tests into its value-added system (reading, math, science, social studies, and language arts). New York City strictly relies on the state's math and English language arts exams. Neither the Texas nor the New York state test was designed on a continuous developmental (or "vertically equated") scale.

Is the measurement tool appropriate?

In assessing a broad set of skills, an instrument must be devised that provides a valid and reliable inference about students' mastery of those skills. No test will cover all of the standards that students are expected to master. By necessity, a test instrument must sample items from a much broader domain of skills. Only by drawing an even and representative sample from this broader domain can a test provide a valid inference about student learning in that domain.

However, such tests are the exception, not the rule. Many skills simply are not amenable to standardized tests and, inevitably, are underrepresented on the test. Many skills that can be tested never appear on the test. Others are over-represented on the test. Teachers aware of systematic omissions and repetitions can substantially inflate scores by narrowly focusing on these items or by "teaching to the format" of the test. Recent studies of the New York, Texas, and Massachusetts tests find that some parts of the state curriculum never appear on the test (Jennings & Bearak 2010; Holcombe, Jennings & Koretz 2010). For example, 50 percent of the possible points on the 2009 New York eighth-grade math test were based on only seven of the forty-eight state standards; only a score of 51 percent was required to pass.

A useful way to look at the importance of the test itself is to compare value-added calculations from more than one test. Since 1998, Houston has administered two standardized tests annually: the Texas Assessment of Knowledge and Skills (TAKS) and the nationally normed Stanford Achievement Test. Using Houston data, I calculated separate value-added measures for fourth- and fifth-grade teachers on the two tests in the same subject,

using the same students, tested at approximately the same time of year. Teachers who had high value-added on one test tended to score well on the other, but there were many inconsistencies. Many teachers who scored in the top category of the TAKS reading test ranked among the lowest categories on the Stanford test, and vice versa.

In a related study, Papay (2010) calculated ASPIRE bonuses using value-added estimates from separate tests and found that "simply switching the outcome measure would affect the performance bonus for nearly half of all teachers and the average teacher's salary would change by more than $2,000" (p. 3). Such wild inconsistencies certainly run counter to the intended goals of value-added assessment.

Can a teacher's unique effect be isolated?

The successful use of value-added requires a high level of confidence in the attribution of achievement gains to specific teachers. One must be confident that other explanations for test score gains have been accounted for before rewarding or punishing teachers based on these measures. In practice, there are a countless number of factors that hinder our ability to isolate a teacher's unique effect on achievement.

Given one year of test score gains, it is impossible to distinguish between the teacher's effect and other classroom-level factors. Over many years, unusual swings average out, making it easier to infer teachers' own effects, but this is of little comfort to a teacher or school leader looking for actionable information today. What is more, teachers with the fewest years of data – novice teachers – arguably have the most to gain from feedback on their performance. Yet the value-added scores for these teachers are the least reliable.

Most value-added systems in practice – including New York City's – fail to separate teachers' influences from school-level effects on achievement. But performance differs systematically across schools due to differences in school policy, leadership, discipline, staff quality, and student mix. Recent research suggests that school factors can and do affect teachers' value-added. Jackson and Bruegmann (2009) found that students perform better when their teachers have had more effective colleagues. Other studies have found effects of principal leadership on student outcomes (Clark, Martorell & Rockoff 2009). Consequently, teachers rewarded or punished for value-added may be rewarded or punished, in part, based on the colleagues with whom they work.

Who counts?

Value-added systems, in practice, ignore a large fraction of the educational enterprise. Only a minority of teachers teach subjects amenable to standardized testing; not all students are tested; and not all tested students contribute to value-added scores. From the standpoint of value-added assessment, these students and teachers do not count.

In most states, students are tested in reading and math in grades three to eight and again in high school. Other subjects, including science and social studies, are tested less frequently. Because value-added requires last year's test score, only teachers of reading and math in grades four to eight are typically assessed using value-added. Thus, elementary, middle school, and high school teachers of all subjects other than reading and math are ignored by value-added assessment.

Some students are routinely exempted from testing or, for one reason or another, are missing a test score. Large urban districts often have a large number of these cases. I examined data from Houston to see how missing data can affect "who counts" toward a teacher's value-added assessment. I looked at the percentage of students in grades four to six over eight years of testing who were tested in two consecutive years and thus can contribute to a value-added score. Because of disabilities, limited English ability, absenteeism, and other reasons, roughly 14 percent of students in Houston lack a test score in any given year. As many as 16 percent of Black students lack scores, and close to 30 percent of recent immigrants are not tested.

The percentage of students who have both a current and prior year test score is even lower. Only 66 percent of all students had both scores, a fraction that falls to 62 percent for Black students, 47 percent for ELL students, and 41 percent for recent immigrants. Thus, in a given year, depending on the group, 40 percent to 60 percent of students in this population do not count toward teachers' value-added assessments.

This issue is more than just a technical nuisance. To the extent that districts reward or punish teachers on the basis of value-added, they risk ignoring teachers' efforts with a substantial share of their students and provide no incentive for teachers to invest in students who will not count. Unfortunately, districts like New York City and Houston have very large numbers of mobile, routinely exempted, and frequently absent students, and these students are unevenly distributed across schools and classrooms. Teachers serving these students in disproportionate numbers are most likely to be affected by a value-added system that – by necessity – ignores many of their students.

Are value-added scores precise enough to be useful?

Some uncertainty is inevitable in value-added measurement, but for practical purposes it is worth asking: Are value-added measures precise enough to be useful in high-stakes decision-making or for professional development? Using the example given earlier, a teacher ranked in the 43rd percentile on New York City's Teacher Data Report might have a range of possible scores from the 15th to the 71st percentile after taking statistical uncertainty into account. What is the source of this imprecision? Recall that value-added measures are estimates of a teacher's contribution to student test-score gains. The more certain we can be that gains are attributable to a specific teacher, the more precise our estimates will be. The best way to improve this certainty is to have more years of classroom test results. Thus, experienced teachers will tend to have more precise estimates than new teachers.

To get a better sense of the average level of uncertainty in New York City's Teacher Data Reports, I examined the full set of value-added estimates reported by that system in 2008-2009. As expected, the level of uncertainty is higher when only one year of test results is used versus three. But in both cases, the average range of percentiles is very wide. For example, in math (and using all years of available data, which provides the most precise possible measures), the average range is about 34 per-

centage points (e.g., from the 46th to 80th percentile). When looking at only one year of test results, the average range increases to 61 percentage points (e.g., from the 30th to the 91st percentile).

The average level of uncertainty is higher still in English language arts and in sections of the city with high levels of student mobility, such as the Bronx. Given the level of uncertainty reported in the data reports, half of all teachers in grades four to eight have wide enough performance ranges that they cannot be statistically distinguished from 60 percent or more of all other teachers in the city.

Using New York City's performance categories, we cannot rule out the possibility that a teacher with a range of percentiles from 15 to 71 is "below average," "average," or close to "above average." It is unclear what this teacher or his principal can do with this information to improve instruction or raise student performance. More years of data help, but the promise that better data will be available in the future is of little use to a teacher looking for guidance in real time. Value-added results for student subgroups might hold greater promise, to the extent that they highlight areas in need of improvement. Yet in most cases, the number of students used to calculate these subgroup scores is so small that the resulting level of uncertainty renders them meaningless.

It is interesting to point out that, by definition, 50 percent of teachers will perennially fall in the "average" performance category on New York City's Teacher Data Report. Another 40 percent will be considered "below average" or "above average." The remaining 10 percent are either exceptional (top 5 percent) or failing (bottom 5 percent). Thus, out of all teachers issued a value-added report each year, half will be told little more than that they are "average." At most, one in three will receive a signal that improvement is needed, though high levels of uncertainty will raise some doubt about this signal. In no case will teachers be told what actions need to be taken. Of course, teachers persistently in the top 5 percent are almost certainly worth recognizing; teachers persistently in the bottom 5 percent deserve immediate scrutiny. Still, it seems a great deal of effort has been expended to identify a very small fraction of teachers. In the end, a tool designed for differentiating teacher effectiveness has done very little of the sort.

Is value-added stable from year to year?

Given the extent of uncertainty in teacher value-added scores, it would not be surprising if these estimates fluctuated a great deal from year to year. In fact, this is generally what is observed in both Houston and New York City. In Houston, among those in the lowest 20 percent of value-added, only 36 percent remain among the lowest performers in the following year. Similarly, among those in the top 20 percent, only 38 percent remain among the top performers the next year. Twenty-three percent of last year's lowest performers are among the top performers in the following year, and vice versa. A similar pattern holds in an analysis of New York City Teacher Data Report data.

Again, imprecision and variability is reduced as additional years of classroom data accumulate.

But here again, this knowledge is of little use in real time. A top-performing teacher may be awarded (or punished) one year based on her latest round of test results, only to get the opposite feedback the following year. Wisely, districts that have adopted value-added systems – including New York City – caution users against making rash decisions based on one year of estimates. But, this estimate is one of only a few made available in value-added assessment systems. Inexperienced teachers – those most in need of immediate feedback – simply will not have the multiple years of data on which to rely. It seems unlikely that teachers and their school leaders will not pay close attention to these noisy and imprecise estimates.

Discussion

In the abstract, value-added assessment of teacher effectiveness has great potential to improve instruction and, ultimately, student achievement. The notion that a statistical model might be able to isolate each teacher's unique contribution to his or her students' educational outcomes – and by extension, their life chances – is a powerful one. With such information in hand, one could not only devise systems that reward teachers with demonstrated records of success in the classroom – and remove teachers who do not – but also create a school climate in which teachers and principals work constructively with their test results to make positive instructional and organizational changes.

But the promise that value-added systems can provide such a precise, meaningful, and comprehensive picture is much overblown. As this report argues, value-added assessments – like those reported in the New York City Teacher Data Reports and used to pay out bonuses in Houston's ASPIRE program – are, at best, a crude indicator of the contribution that teachers make to their students' academic outcomes. Moreover, the set of skills that can be adequately assessed in a manner appropriate for decisions based on value-added represents a small fraction of the goals our nation has set for our students and schools.

The implementation of value-added systems faces many challenges. Not all students are tested, and many, if not a majority of teachers do not teach tested subjects. Students without a prior-year test score – such as chronically mobile students, exempted students, and those absent on the day of the test – simply do not count toward teachers' value-added estimates. In many districts, these students constitute a substantial share of many teachers' classrooms.

Often, state tests are predictable in both content and format, and value-added rankings will tend to reward those who take the time to master the predictability of the test. Evidence from Houston presented here showed that one's perception of a teacher's value-added can depend heavily on which test one looks at. Annual value-added estimates are highly variable from year to year and, in practice, many teachers cannot be statistically distinguished from the majority of their peers. Persistently exceptional or failing teachers – say, those in the top or bottom 5 percent – may be successfully identified through value-added scores, but it seems unlikely that school leaders would not

already be aware of these teachers' persistent successes or failures.

Research on value-added remains in its infancy, and it is likely that these methods – and the tests on which they are based – will continue to improve over time. The simple fact that teachers and principals are receiving regular and timely feedback on their students' achievement is an accomplishment in and of itself, and it is hard to argue that stimulating conversation around improving student achievement is not a positive thing. But teachers, policy-makers, and school leaders should not be seduced by the elegant simplicity of "value-added." Before adopting these measures wholesale, policy-makers should be fully aware of their limitations and consider whether the minimal benefits of their adoption outweigh the cost.

References

Clark, Damon, Paco Martorell, and Jonah Rockoff. 2009. "School Principals and School Performance." CALDER Working Paper No. 38. Washington, DC: Urban Institute.

Holcombe, Rebecca, Jennifer L. Jennings, and Daniel Koretz. 2010. "Predictable Patterns that Facilitate Score Inflation: A Comparison of New York and Massachusetts." Working Paper. Cambridge, MA: Harvard University.

Jackson, C. Kirabo, and Elias Bruegmann. 2009. "Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers," *American Economic Journal: Applied Economics* 1:85–108.

Jennings, Jennifer L., and Jonathan M. Bearak. 2010. "Do Educators Teach to the Test?" Paper presented at the Annual Meeting of the American Sociological Association, Atlanta.

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* 27:615–631.

Papay, John P. 2010. "Different Tests, Different Answers: The Stability of Teacher Value-Added Estimates Across Outcome Measures," *American Education Research Journal,* published online (April 19); print version forthcoming.

Annenberg
Institute for
School Reform

AT BROWN UNIVERSITY

Providence

Brown University
Box 1985
Providence, RI 02912
T 401.863.7990
F 401.863.1290

New York

233 Broadway, Suite 720
New York, NY 10279
T 212.328.9290
F 212.964.1057

www.annenberginstitute.org