EDUCATION POLICY FOR ACTION SERIES

EDUCATION CHALLENGES FACING NEW YORK CITY

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

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

CONTENTS

FIGURES

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.

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 adopted these methods en masse in recent years in an attempt to objectively quantify teaching effectiveness and promote and retain teachers with a demonstrated record of success.

Attention to the quality of the teaching force makes a great deal of sense. No other school resource is so directly and intensely focused on student learning, and research has found that teachers can and do vary widely in their effectiveness (e.g., Rivkin, Hanushek & Kain 2005; Nye, Konstantopoulos & Hedges 2004; Kane, Rockoff & Staiger 2008).[1] Furthermore, teacher quality has been found to vary across schools in a way that systematically disadvantages poor, low-achieving, and racially isolated schools (e.g., Clotfelter, Ladd & Vigdor 2005; Lankford, Loeb & Wyckoff 2002; Boyd et al. 2008).

But questions remain as to whether value-added measures are a valid and appropriate tool for identifying and enhancing teacher effectiveness. In this report, I aim to provide an accessible introduction to these new measures of teaching quality and put them into the broader context of concerns over school quality and achievement gaps. Using New York City's Teacher Data Initiative and Houston's ASPIRE (Accelerating Student Progress, Increasing Results & Expectations) program as case studies, I assess the potential for these measures to improve outcomes in urban school systems. In doing so, I outline some of the most important challenges facing value-added measures in practice.

1 Value-Added Measurement: Motivation and Context

Traditional measures of teacher quality have always been closely linked with those found in teacher pay schedules: years of experience, professional certification, and degree attainment. As recently as the 2001 No Child Left Behind Act (NCLB), teacher quality was commonly formalized as a set of minimum qualifications. Under NCLB, "highly qualified" teachers of core subjects were defined as those with at least a bachelor's degree, a state license, and demonstrated competency in the subject matter taught (e.g., through a relevant college major or master's degree).

However, these minimum qualifications have not been found by researchers to be strongly predictive of student outcomes on standardized tests (e.g., Goldhaber 2008; Hanushek & Rivkin 2006; Kane, Rockoff & Staiger 2008). Knowing that a teacher possesses a teaching certificate, a master's degree, or a relevant college major often tells us little about that teacher's likelihood of success in the classroom. There are many reasons not to totally dismiss

[1] This literature is frequently misinterpreted as stating that teacher quality is more important for student achievement than any other factor, including family background. Statements such as "Studies show that teachers are the single most important factor determining students' success in school" have appeared in dozens of press releases and publications in recent years. For an example, see the May 4, 2010 statement from the U.S. House Committee on Education and Labor at <http://edlabor.house.gov/newsroom/2010/05/congress-needs-to-support-teac.shtml>. I know of no study that demonstrates this.

these qualifications,[2] but common sense suggests this information alone can only be a crude indicator for differentiating teaching effectiveness.

Over the past fifteen years, research on teacher quality has adopted a new paradigm: measuring effectiveness on the basis of *student outcomes*, as opposed to *teacher inputs* (e.g., Rivkin, Hanushek & Kain 2005; Rockoff 2004; Sanders & Horn 1994). While outcome-based measures of teaching effectiveness are not a new concept (Murnane & Cohen 1986; Odden & Kelley 2002), several forces have converged to reorient the definition of teacher quality around student achievement. First, recent policies of high-stakes accountability have increased pressure on schools to measure and demonstrate results. Given teachers' close contact with students, the extension of high-stakes accountability to individual teachers was perhaps inevitable. Second, new longitudinal data systems now exist that allow student achievement to be tracked over time and matched to classroom teachers.[3] Arguably, no credible test-score–based system of teacher evaluation could exist in the absence of such systems. Third, advancements in data-processing capacity and statistical modeling have yielded an array of value-added techniques with the potential for isolating teachers' unique contribution to student outcomes.

Perhaps most importantly, political and philanthropic preferences have aligned to bring about seismic shifts in our conception of teaching effectiveness. Leaders of both political parties have strongly endorsed linking teacher evaluation to student test scores, and foundations such as the Bill & Melinda Gates Foundation, the Milken Family Foundation, and the Broad Foundation have provided significant financial resources to support these efforts. In promoting Race to the Top – President Barack Obama's $4 billion competitive grant program aimed at systemic education reform – President Obama (2009) stated, "Success should be measured by results. . . . That's why any state that makes it unlawful to link student progress to teacher evaluation will have to change its ways." It is widely believed that Race to the Top will serve as a template for the reauthorization of the federal Elementary and Secondary Education Act in coming years.

Race to the Top is quite specific in its conception of an "effective teacher" (see Appendix A). "Highly effective teachers" are those whose students achieve high rates of growth, defined – narrowly – by the program as a change in test scores between two or more points in time (U.S. Department of Education 2010). Supplemental measures are encouraged, as are alternative metrics for teachers of non-tested grades and subjects, but the primary emphasis rests

[2] For example, these qualifications can be viewed as minimum expectations for classroom teachers. Evidence that some practicing teachers without these qualifications are as effective as those who possess them cannot tell us how the wholesale removal of these minimum requirements would affect the quality of the teaching force. Moreover, virtually all of the research on teacher qualifications has focused on a narrow measure of success: single-year gains on standardized tests. We know little about how qualifications affect other outcomes, such as organizational stability, student behavior and motivation, aspirations, engagement, persistence, and the like.

[3] The Data Quality Campaign, founded by the Bill & Melinda Gates Foundation, tracks and supports state efforts to create these systems; see <www.dataqualitycampaign.org>.

squarely on test-score growth. One of the program's major selection criteria, "Great Teachers and Leaders," contributes at least 70 of the 500 possible application points to the linking of teacher evaluation and student test performance. For example, in their applications, states will be judged by the extent to which they or their districts (U.S. Department of Education 2010):

- measure individual student growth;
- implement evaluation systems that use student growth as a significant factor in evaluating teachers and principals;
- include student growth in annual evaluations;
- use these evaluations to inform professional support, compensation, promotion, retention, tenure, and dismissal;
- link student growth to in-state teacher preparation and credentialing programs, for public reporting purposes and the expansion of effective programs;
- incorporate data on student growth into professional development, coaching, and planning.

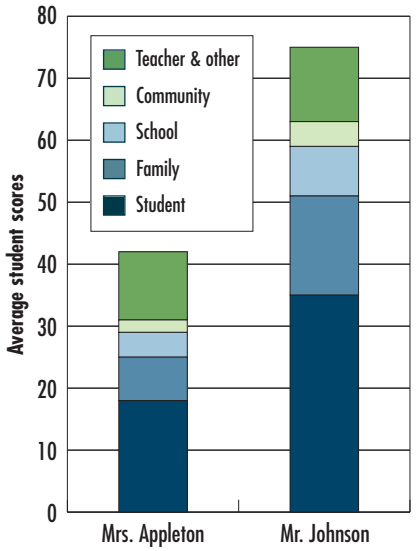
Race to the Top and the "new view" of teacher effectiveness have stimulated a largely productive and long-overdue discussion between policy-makers, researchers, and the public over how to assess teacher quality and address, develop, and support under-performing teachers. And it is fair to say that there is little enthusiasm for the traditional model of teacher evaluation used in many public schools: infrequent classroom observations and a pro forma tenure (Toch & Rothman 2008; Weisburg et al. 2009). But whether or not the shift to intensive use of value-added measures of effectiveness will improve our nation's system of teaching and learning remains to be seen. Indeed, there are good reasons to believe these measures may be counterproductive.

2 What Is a Teacher's Value-Added?

It is useful to distinguish between the theoretical conception of value-added and the methods used to calculate it for an actual set of teachers.[4] In theory, a teacher's value-added is the unique contribution she makes to her students' academic progress. That is, it is the portion of her students' success (or lack thereof) that cannot be attributed to any other current or past student, school, family, or community influence. Any system of test-based teacher accountability that does not have value-added as its ultimate object of interest risks crediting or faulting teachers for outcomes beyond their control.

Despite its theoretical appeal, isolating a teacher's unique contribution is a very difficult exercise. A simple example will help illustrate this point. Suppose recent budget cuts have forced a district to lay off one of its two fourth-grade teachers, and the district prefers to dismiss the least effective of the two, as evidenced by their state test results. Mrs. Appleton's students averaged a 42 on the most recent math exam, while Mr. Johnson's students averaged a 75 (see Figure 1).

Figure 1
Factors affecting average achievement in two classrooms: hypothetical decomposition



Is it fair to say that Mr. Johnson is the more effective teacher? Not without more information. Mr. Johnson's higher scores could reflect a host of factors that have nothing to do with his effectiveness in the classroom: greater family resources and involvement, higher initial levels of student ability, superior third-grade instruction, greater out-of-school support, and so on. The hypothetical contributions of these other factors to average achievement are illustrated by the colored bars in Figure 1. One could look for a way to statistically "remove" the effects of these influences from the achievement measure, but many of the factors that matter most – parental support and student ability, for example – are difficult, if not impossible, to quantify.

An alternative comparison is the extent of student progress from year to year (see Figure 2). If we assume that many of the external factors influencing a student's fourth-grade achievement are the same as those influencing her third-grade achievement, then the change in the student's score will cancel out these effects and reveal only the impact of changes since the third-grade test, with the year of fourth-grade instruction being the most obvious.[5] Importantly, the focus on gains takes into account

[4] There are many excellent and readable introductions to value-added methods. In writing this section, I benefited greatly from Braun (2005), Buddin et al. (2007), Koretz (2008), Rivkin (2007), Harris (2009), and Hill (2009).

[5] This statement assumes something about the scale on which achievement is measured. I return to this point in sections 4 and 5.

that these teachers' students began the year at very different levels. The idea is illustrated in Figure 2: Mrs. Appleton's students started out at a much different point than Mr. Johnson's, but most of the factors determining these initial differences in achievement "net out" in the average gain score. The remainder represents the impact of the fourth-grade teacher, as well as other influences that may have produced differential growth between the two tests.

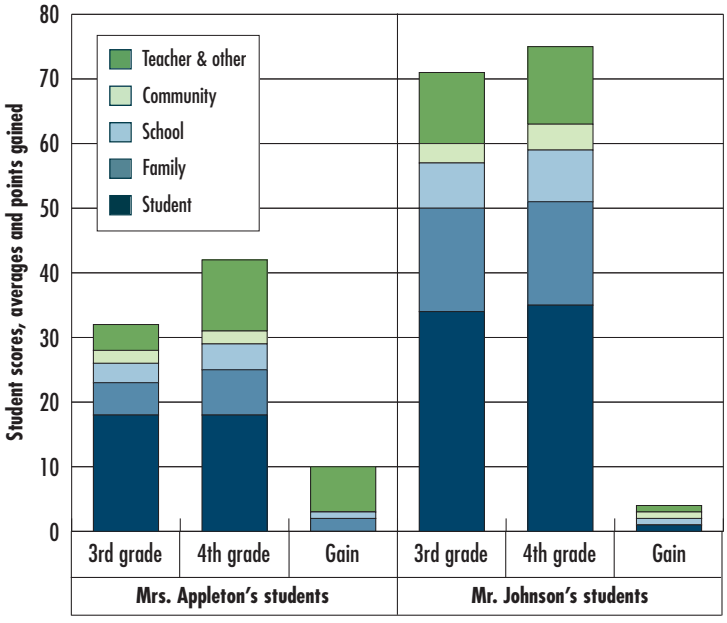
Figure 2 shows that Mrs. Appleton's students had an average gain of ten points, while Mr. Johnson's gained an average of four. Can we now declare Mrs. Appleton the more effective math teacher? Do these gain scores represent these teachers' value-added? Not necessarily. While we may have removed the effects of fixed differences between student populations, we need to be confident that we have accounted for other factors that contributed to changes in test performance from third to fourth grade. These factors are potentially numerous: family events, school-level interventions, the influence of past teachers on knowledge of this year's tested material, or a disruptive or especially helpful student in the class, among others. Many of these factors are random events, while others systematically affect teachers from year to year.

If value-added measures are to be successfully used in practice to recognize effective teachers, one needs a high level of confidence in the attribution of achievement gains to specific teachers. Were students randomly assigned to teachers, this would be straightforward: any systematic differences between classroom achievement gains would almost certainly be due to the teacher. All other factors influencing year-to-year changes would effectively average out, allowing us to detect real differences in achievement gains across teachers. In reality, students are not randomly assigned to classes – in many cases, quite purposefully so. Consequently, value-added methods use a statistical model to answer the question: "How would these students have fared if they had not had [Mrs. Appleton or Mr. Johnson] as a teacher?"

This is a difficult question that is taken up in the next section. For now, it is useful to think of a teacher's value-added as her students' average test-score gain, "properly adjusted" for other influences on achievement. The New York City Teacher Data Initiative (TDI) and the Houston ASPIRE programs are two prominent value-added systems that have their own methods for "properly adjusting" student test scores. These two programs and their methods are described in the next section.

Figure 2

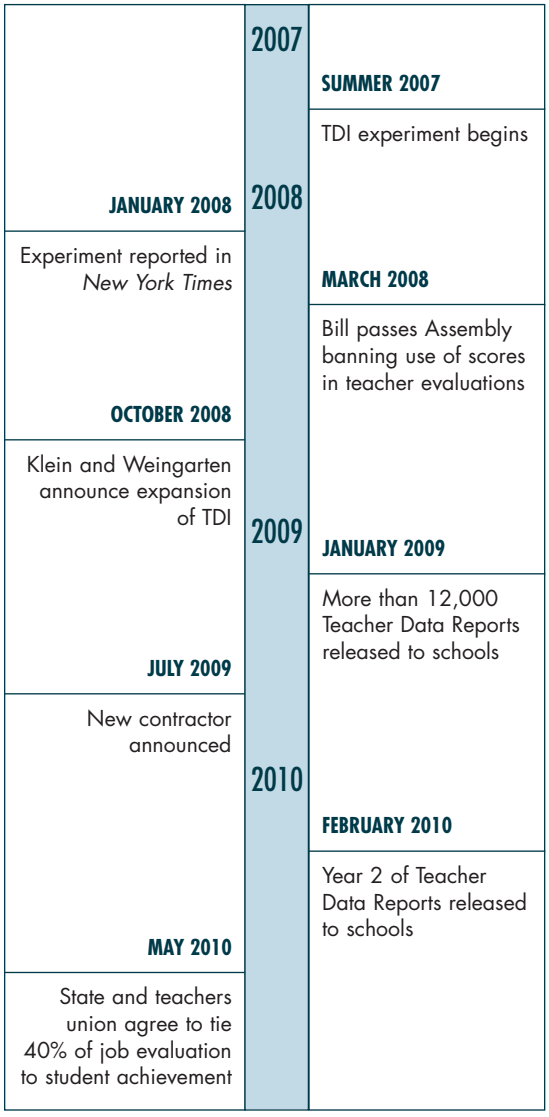
Factors affecting year-to-year test score gains in two classrooms: hypothetical decomposition



3 Value-Added in Practice: New York City and Houston

The New York City Department of Education (NYCDOE) and Houston Independent School District (HISD) have been at the forefront of developing and implementing value-added measures of teacher effectiveness in their districts. Other large school districts, including Dallas, Denver, Minneapolis, and Washington, D.C., have made extensive use of value-added methods in rewarding and evaluating teacher effectiveness. Although the New York and Houston programs have similar goals, they have quite different histories and methodologies, as described in this section.

Figure 3
New York City Teacher Data Initiative timeline



The New York City Teacher Data Initiative

In January 2008, the *New York Times* revealed that the NYCDOE was conducting an experiment with a team of academic researchers that randomly distributed teacher value-added reports to 140 participating school principals and collected subjective teacher evaluations from these and an additional 140 control principals (Medina 2008b; Rockoff et al. 2010). The experiment was intended to reveal whether and how principals use teacher value-added reports in practice. At the time, the NYCDOE publicly supported the idea of using test-score–based measures of teacher performance, but had no official position on how these measures should be used (Medina 2008b; Keller 2008).[6]

Two months later – possibly in response to the *New York Times* revelation – the New York State Assembly passed a controversial bill that barred New York City and other districts in the state from tying teacher tenure decisions to student test scores (Medina 2008a) (see Figure 3 for a timeline of these events). The bill – supported by then-president of the United Federation of Teachers (UFT) Randi Weingarten – was later signed into law by Governor David Patterson, intended to be in effect through June 2010.

Despite the official ban on using test scores for tenure decisions, the NYCDOE pressed forward with the TDI, with conditional support from Weingarten and the UFT. In accordance with the law, TDI information was explicitly not to be used for rewarding or dismissing teachers. As stated in a joint Klein/Weingarten letter to teachers in 2008, the TDI's value-added reports were to be used solely for professional development, to "help [teachers] pinpoint [their] own strengths and weaknesses, and . . . devise strategies to improve" (Klein & Weingarten 2008).

The TDI released its first complete set of Teacher Data Reports to more than 12,000 teachers in 2009. These reports consisted of separate analyses of English language arts (ELA) and mathematics test results and were generated for teachers who had taught these subjects in grades four to eight in the prior year. (A more detailed description of the report itself is provided later in this section.) A second year of reports was released in 2010, reflecting significant revisions made by the NYCDOE and its new contractor, the Wisconsin Value-Added Research Center.[7]

The NYCDOE has expressed several broad goals for the TDI program.[8] First, its data reports are intended to provide measures of value-added that can be reported to principals and teachers in an accessible and usable form. The reports are to be viewed as "one lens" on teacher quality that should be "triangulated" with other information about classroom effectiveness to improve performance. The reports are seen as an "evolving tool" that will continue to be refined over time based on principal and teacher feedback. Second, it is hoped the reports will "stimulate conversation" about student achievement within schools and promote better instructional practices through professional development. Finally, the measures will help the district learn more about "what works" in the classroom. Value-added measures have already enabled a wide range of studies on teacher effectiveness in New York City (e.g., Boyd et al. 2009; Kane, Rockoff & Staiger 2008; Rockoff 2008), and the introduction of these measures into schools will enable additional research. In 2009, the NYCDOE and UFT signed on to participate in a historic, large-scale Gates Foundation study – the Measures of Effective Teaching, or "MET" project – that intends to benchmark value-added measures against alternative measures of teaching effectiveness and identify practices that are associated with high value-added (Medina 2009).[9]

New York's law banning the use of test scores for teacher evaluation would later complicate the state's application for Race to the Top funding. Guidelines for the federal grant program explicitly penalized states with such laws, and Mayor Bloomberg and members of the state legislature pushed for a reversal. Speaking in Washington in November 2009, Mayor

[6] In another article appearing that month, then–Deputy Chancellor Chris Cerf stated that he was "unapologetic that test scores must be a central component of evaluation" (Keller 2008).

[7] See <http://varc.wceruw.org/>.

[8] This description is compiled from a phone interview with Amy McIntosh, Joanna Cannon, and Ann Forte of the NYCDOE (January 14, 2010), an October 2008 presentation by Deputy Chancellor Chris Cerf ("NYC Teacher Data Initiative"), and a September 2008 training presentation by Martha Madeira ("NYC Value-Added Data for Teachers Initiative").

[9] For information on the MET project, see <http://metproject.org/project>.

Bloomberg announced that he had "instructed City Schools Chancellor Joel I. Klein to begin using student performance data immediately to inform teacher tenure decisions."[10] The mayor further encouraged the state legislature to mandate the use of student performance data in teacher evaluation systems statewide. The law was not repealed in time for the state's first bid for Race to the Top, and New York ultimately failed to receive an award in the first round.

For its second Race to the Top bid in summer 2010, New York took dramatic steps toward tying teacher evaluations to student progress. An agreement between the state department of education and the teachers' unions linked 40 percent of teachers' performance evaluations to student achievement measures (Medina 2010), though these measures were not based on value-added alone. The agreement did not go as far as some states – including Florida, Indiana, Rhode Island, Tennessee, and Colorado – that chose to base more than half of teachers' job evaluations on student performance, but it was a major departure from the 2008 law banning the practice altogether.[11] In 2010, the U.S. Department of Education announced New York as one of the Race to the Top winners, awarding the state $700 million in federal aid.

The results of the 2008 NYCDOE value-added experiment were finally released to the public in July 2010 (Rockoff et al. 2010). Authors of the study found that value-added measures on the Teacher Data Reports were positively related to principals' subjective ratings of teachers collected prior to their knowledge of the test results. That is, teachers deemed more effective based on value-added were more likely to have been rated as effective by their principals. More importantly, however, principals randomly selected to receive Teacher Data Reports changed their evaluations of teachers in response to the new information. Less effective teachers were more likely to leave their schools or be assigned "unsatisfactory" ratings when their principal received a value-added report.

Early indications – such as the value-added experiment described above – suggest that principals can and will use value-added information in their assessment of teacher effectiveness. More recent data released by the NYCDOE showed that a higher fraction of low–value-added teachers were denied tenure in 2010 than were high–value-added teachers (Martinez 2010). If principals are to use value-added reports in making consequential personnel decisions, it is critical that their underlying measures be valid, reliable, and precise indicators of teacher effectiveness (criteria that are described in Section 4). Just as importantly, users must have a rich understanding of their methods and limitations.

Appendix B illustrates a sample New York City Teacher Data Report from 2010 produced by the NYCDOE for "Mark Jones," a fictional eighth-grade teacher at "I.S. 000." This particular report is for mathematics; if Mr. Jones also taught ELA, he would receive a separate report for that subject.

The Teacher Data Report provides value-added measures summarized in a number of ways. (The calculation itself is explained later in this section.) The most important thing to note is that the key metric on the report is the teacher's value-added percentile in the citywide distribution of teachers teaching the same grade and subject, with similar amounts of

experience (in Mr. Jones's case, ten years). That is, results are not reported in units of "achievement," but rather as percentile rankings.[12]

Thus, value-added, in practice, is a *relative* concept. Teachers are, in effect, graded on a curve – a feature that is not always obvious to most observers. A district with uniformly declining test scores will still have "high" and "low" value-added teachers; a district's logical aspiration to have exclusively "high value-added" teachers is a technical impossibility. The value-added percentile simply indicates where a teacher fell in the distribution of (adjusted) student test-score gains.

Value-added is reported for both last year's test results (in this case, 2008-2009) and on all prior year's test results for that teacher (in this example, the last four years). Mr. Jones's value-added places him in the 43rd percentile among eighth-grade math teachers last year; that is, 43 percent of teachers had lower value-added than he did (and 57 percent had higher value-added). His value-added based on the last four years of results places him in the 56th percentile. The percentiles are then mapped to one of five performance categories: "high" (above the 95th percentile), "above average" (75th to 95th), "average" (25th to 75th), "below average" (5th to 25th), and "low" (below 5th). Mr. Jones's percentile rankings would appear to place him squarely in the "average" performance category.

Another element of the Teacher Data Report worth noting is the reported range of percentiles associated with Mr. Jones's value-added ranking (the black line extending in two directions from his score). In statistical terminology, this range is referred to as a "confidence interval." It represents the level of uncertainty associated with the value-added percentile measure. As the report's instructions describe these ranges: "We can be 95 percent certain that this teacher's result is somewhere on this line, most likely towards the center." These ranges – or confidence intervals – are discussed more in Section 4. For now, note that Mr. Jones's range for the prior year's test extends from (roughly) the 15th percentile to the 71st. Based on his last four years, his range extends from the 32nd percentile to the 80th. His value-added percentiles – 43 and 56 – fall in the middle of these ranges.

On the second page of the data report (see <http:/schools.nyc.gov>), value-added measures and percentiles are reported for several subgroups of students: initially high-, middle-, and low-achieving students (based on their prior year's math achievement); boys and girls; English language learners; and special education students. Mr. Jones performed at the "above average" level with his initially high-achieving students, but fell into the "average" category for all other subgroups. Ranges, or confidence intervals, are also reported for each of these subgroups.

How are these value-added percentiles calculated, exactly?[13] Recall that a teacher's value-added can be thought of as her students' aver-

[10] Press release PR-510-09, Office of the Mayor, November 25, 2009.

[11] On Colorado and Tennessee, see "Colorado Approves Teacher Tenure Law," *Education Week*, May 21, 2010, and "Tennessee Lawmakers Approve Teacher Evaluation Plan," *Memphis Commercial Appeal*, January 15, 2010.

[12] I address the reported "proficiency" scores (e.g., 3.27, 3.29) later in the report.

[13] A useful and concise explanation is provided at the top of the Teacher Data Report itself. We benefited from a more technical explanation of the 2009 methodology in an internal 2009 technical report by the Battelle Memorial Institute. The model is similar to that estimated by Gordon, Kane, and Staiger (2006) and Kane, Rockoff, and Staiger (2008).

age test scores, "properly adjusted." New York City operationalizes this idea by comparing students' actual scores under a given teacher to their predicted score. This predicted score can be thought of as each student's counterfactual level of achievement – that is, their predicted achievement had they been taught by a different teacher (say, the average teacher in the district). The prediction itself is based on a number of things, the most important of which is the student's prior achievement. How a student actually performed under Mrs. Appleton relative to how he would have performed under a different teacher represents Mrs. Appleton's value-added for that student.

A stylized example will help specify this idea. Suppose students are given a math test each year that is scored between 1 and 100. Two of Mr. Johnson's fourth-grade students, Melissa and Doug, earned a 42 and a 65 on the test, respectively. As noted in Section 2, there are a host of possible explanations for their performance on this test, among which Mr. Johnson's instruction is but one. For example, Melissa was identified last year with a learning disability, and her low third-grade math score – 35 on the same scale – reflected this. Doug scored relatively high on his third-grade test, earning a 72. Melissa comes from a single-parent family whose low income qualifies her for reduced-price lunches. Doug, on the other hand, has two upper-middle-class parents.

Figure 4
New York City value-added model: predictors

Student characteristics	Prior year achievement ELA Prior year achievement math Free or reduced-price lunch eligibility Special education status English learner status Number of suspensions, absences Student retained in grade Student attended summer school Student new to school Race Gender
Classroom characteristics	Classroom averages of each of the above Class size
School characteristics	Average classroom characteristics Average class size

Source: New York City Department of Education. Based on the 2009 Teacher Data Reports.

Because the school district has richly detailed data on thousands of students' academic histories, it can provide a statistical prediction of how Melissa and Doug were likely to perform on their fourth-grade math test. Effectively, this prediction is the average district score for students with Melissa and Doug's prior achievement and other relevant characteristics. (The full list of student, classroom, and school characteristics used in New York City's model is shown in Figure 4.) Suppose the model predicts that Melissa – with her past achievement and other characteristics –would be expected to earn a 39 on her fourth-grade test, while Doug would be expected to score a 76.

The three-point difference between Melissa's actual score of 42 and her predicted score of 39 represents Mr. Johnson's value-added to Melissa's achievement. Similarly, the four-point negative differential between Doug's score of 72 and his predicted score of 76 represents Mr. Johnson's value-added (or "value-subtracted"?) to Doug's achievement. If we repeat this exercise for every student in Mr. Johnson's class, we can average the results and call this his value-added for that year.[14] If Mr. Johnson has taught

for ten years, we could also average his students' value-added measures over all of those years.

A few key features of this approach are worth highlighting. First, students' predicted scores are based on how other students with similar characteristics and past achievement performed – who were taught by other teachers in the district. Thus, value-added is inherently relative: it tells us how teachers measure up when compared with other teachers in the district or state who are teaching similar students. Second, test scores are rarely of the vertical scale type suggested by the above example. That is, we can rarely say that a student like Melissa moved from a 35 to a 42 on the "math" scale as she progressed from third to fourth grade.

As a compromise, most value-added methods rescale test scores to have a mean of zero and a standard deviation (SD) of one.[15] This new scale tells us where students are in the distribution of test scores in each grade. For example, Melissa may have moved from a -0.25 to a -0.20 on this scale, from 0.25 SDs below the average third-grader to 0.20 below the average fourth-grader, a gain of 0.05 SD. Under certain assumptions, this scale is appropriate, but it is important to keep in mind that students, like teachers, are being measured on a curve – that is, relative to other tested students in the same grade.

Rather than reporting results on the admittedly less-than-transparent SD scale, the NYCDOE converts value-added results to a scale with which teachers are familiar: the state's "performance levels" (1 to 4). On Mark Jones's report, we see that his twenty-seven students were predicted to perform at an average of 3.29 in math, somewhere between proficient (3) and advanced (4).[16] In practice, his class averaged 3.26 last year, for a value-added on this scale of -0.03. Over Mr. Jones's past four years, his value-added was +0.03. All of Mr. Jones's subgroup results are presented in the same way, with predicted and actual performance levels, value-added, and a percentile based on this value-added. Of course, the number of students used to estimate each subgroup value-added measure is smaller; for example, Mr. Jones taught eighteen initially high-achieving students and twenty-seven who initially scored in the middle one-third. He also taught more boys (forty-two) than girls (nineteen).

[14] This concept is also known as the year-specific "teacher effect" or "classroom effect" for that teacher and year.

[15] A standard deviation is a measure of variation in a distribution. Loosely, it can be thought of as the "average difference from the mean." For example, the average score on a test might be a 70, with a standard deviation of 8. We can think of this as saying that, on average, students scored 8 points above or below the average of 70. (This is not technically correct, but it is a useful way of thinking about the standard deviation.) The SD depends on the scale of the original measure. Thus we often put measures on a common ("standardized") scale with a mean of zero and SD of one. In this example of a test with a mean of 70 and SD of 8, a student who scored an 82 would receive a score of 1.5 on the alternate scale (1.5 standard deviations above the mean).

[16] New York City's "performance level" scale itself is somewhat puzzling. The performance levels of 1, 2, 3, and 4 are based on cut scores determined by the state at certain points in the test score distribution. They are ordinal categories that represent increasing levels of tested skill (below basic, basic, proficient, and advanced). They are not an interval measure, where the difference in skill between, say, a 2 and a 3 is equivalent to that between a 3 and a 4, nor were they intended to be used in such a way. The NYCDOE further converts its scores to "fractional units" on this scale. For example, a student that receives a raw score that places them between the cut scores of 2 ("basic") and a 3 ("proficient") might be assigned a performance level of 2.42. Because the proficiency categories are ordinal, it isn't clear what a proficiency level of 2.42 means. It plausibly could be interpreted as being 42 percent of the way between basic and proficient, but it remains that a movement of 0.10 between each point on the scale (1 to 4) will represent different gains in achievement. In practice, this unusual system does not present a problem for the Teacher Data Reports. There the value-added measures are calculated using standardized scores and only converted to performance levels after the fact (Source: internal Battelle Memorial Institute technical report, 2009).

Challenges associated with this method of evaluating teachers are discussed in greater detail in Section 4. Generally speaking, the value-added system in place in New York City is no worse and no less valid than models used elsewhere. In fact, in many ways the city's Teacher Data Reports are superior to those found in other districts. The reports themselves are easy to read, and the results presented on the report are explained in clear (and largely accurate) language. The 2010 report was substantially improved and simplified from the 2009 design (pictured in Appendix C), though the underlying concepts and methods remain largely the same. On the back of the report, teachers are encouraged to use their results for self-improvement and are provided a template for thinking about them.

The Houston ASPIRE Program

Unlike the NYCDOE, the Houston Independent School District (HISD) has made its teacher value-added system an explicit part of its performance pay plan since 2007. In recent months, the district moved to allow dismissals based on value-added. This section provides a brief overview of this nationally known program.

ASPIRE, Houston's teacher performance bonus program, is based on value-added calculated using EVAAS (Education Value Added Assessment System), a value-added method pioneered by William Sanders in Tennessee (Sanders, Saxton & Horn 1997). EVAAS is quite different from the model used by the NYCDOE, but its objective is fundamentally the same: isolating a teacher's contribution to student progress.

Houston has awarded bonuses for student performance since 2000, though its awards program was initially tied to aggregate school-level test results (HISD 2009). A new system based on teacher value-added – the Teacher Performance Pay Model – was announced in 2006, with its first set of awards totaling $17 million to be paid in 2007 (Blumenthal 2006). Houston's most recent round of bonuses awarded more than $40 million, up $8.5 million from the prior year (ASPIRE 2010).

The Teacher Performance Pay Model gained substantial notoriety in its first year when bonus payments to teachers and principals were made public in the *Houston Chronicle* and some highly recognized teachers failed to receive awards. Several months later, it was revealed that a computer glitch had overpaid close to 100 teachers (Associated Press 2007). In September 2007, the HISD school board voted to overhaul the system, renaming the program ASPIRE and adopting the EVAAS model (Olson 2007)[17].

ASPIRE consists of three tiers, or "strands." Strand I is a school-level award that rewards all staff in schools where students demonstrated gains and whose progress ranked in the top two quartiles for their grade level. The Strand I progress measure is a school-level value-added measure conceptually similar to that described for New York teachers. Strand II awards individual teachers whose students' progress ranked in the top two quartiles for their grade and

subject. The teacher performance measure is a "value-added cumulative gain index" for a given teacher and subject. Finally, Strand III offers a mix of additional bonus opportunities, including a bonus for attendance.

Taken together, the three strands of ASPIRE amount to a maximum bonus that ranges from $6,600 to $10,300 for classroom teachers. Teachers of self-contained classes can receive awards in Strand II in as many as five subjects: reading, math, language arts, science, and social studies. A teacher scoring in the top quartile in all five subjects can receive a bonus as high as $7,000. According to the *Houston Chronicle*, almost 90 percent of eligible school employees received a bonus for 2008-2009, with classroom teachers earning an average of $3,606 and a maximum of $10,890 (Mellon 2010a).

The EVAAS model is considerably more complex and is much less transparent than the model adopted by the NYCDOE.[18] The model combines results on multiple tests – the Texas Assessment of Knowledge and Skills (TAKS) and the Stanford 10 Achievement Test (or the Aprenda, its Spanish language equivalent) and "layers" multiple years of test results to calculate teachers' cumulative value-added (McCaffrey et al. 2004). Like the New York City system, expected scores in each year are estimated for students in each subject and compared with their actual scores. However, unlike the New York City model, the predicted scores rely on a relatively sparse list of student background characteristics.

In February 2010, the HISD board of education voted to approve the use of value-added measures in teacher tenure decisions (Sawchuk 2010). In a letter to the school board, HISD superintendent Terry Grier stated that his intended use for the value-added measures was for them to be added to "the list of reasons that can be used in teacher dismissal." He expressed a willingness to create "a screening process for principals who propose that teachers gain term contract by requiring them to discuss the performance/effectiveness of all probationary teachers," adding, "this discussion will include the review of value-added. . . . If principals want to grant term contracts to teachers with regressive value-added scores, . . . they should be able to provide a compelling reason for doing so" (Mellon 2010b).

Although ASPIRE's value-added model differs markedly from that used by the Teacher Data Reports in New York City, the programs share the same core objective: differentiating teachers based on their contribution to student achievement and recognizing and rewarding effective teachers. Houston's decision to link its value-added measures explicitly to pay and tenure decisions were precursors to similar decisions in New York City in recent months. In the next section, I provide an overview of the most significant challenges facing value-added measurement in practice, drawing upon data from HISD and the New York City Department of Education to illustrate these challenges.

[17] ASPIRE currently receives funding from several sources: the Broad and Bill & Melinda Gates foundations ($4.5 million), a U.S. Department of Education Teacher Incentive Fund (TIF) grant ($11.7 million), and a Texas District Assessment of Teacher Effectiveness grant. The contract firm Battelle for Kids has provided professional development for the ASPIRE program since 2007.

[18] EVAAS has been sharply criticized for its lack of transparency and inadequate controls for student background. See, for example, Amrein-Beardsley (2008).

4 Challenges to the Practical Implementation of Value-Added

As exemplified by New York City's Teacher Data Reports and Houston's ASPIRE program, value-added measures have been embraced by school leaders nationwide as a means to objectively quantify teacher effectiveness and to reward and retain teachers with a demonstrated record of success. Few can deny these measures' intuitive appeal: if a statistical model can isolate a teacher's unique contribution to students' educational progress, the possibilities for its productive use appear endless. However, these tools have limitations and shortcomings that are not always immediately apparent. Before adopting these measures wholesale, policymakers should be fully aware of their limitations and consider whether the benefits of their adoption outweigh the cost.

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

- 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?

Testable skills

Value-added measurement works best when students receive a single objective numeric test score on a continuous developmental scale – that is, one that is not tied to grade-specific content. The domain of skills that can be adequately assessed in this way is, however, remarkably small. For example, elementary math skill may progress in a way that lends itself to annual standardized testing and a "vertically equated" scale that spans multiple grades. Its basic computational and problem-solving skills are relatively easy to assess on a well-designed short-answer or multiple-choice test.

But history, civics, English literature, music, foreign language, critical thinking, writing, and research skills may not be so easy to assess in this way, and it makes little educational sense to force such skills to conform to such a structure purely for value-added assessment. For this reason, skills readily assessed by standardized tests reflect only a small fraction of what students are expected to know and do. Not all subjects are or can be tested, and even within tested subject areas, only certain skills readily conform to standardized testing. These points are made so frequently that they have virtually lost all meaning; we simply shrug and acknowledge that of course, tests don't capture everything. Yet value-added measures of teaching effectiveness rest exclusively on skills assessable on very narrow standardized tests.

In a recent essay, economist Alan Blinder (2009) persuasively argued that the skills vital for success in the labor market in the near future will be those least amenable to standard-

ized testing: "skills that a computer cannot replicate and that an educated worker in a low-wage country will have a hard time doing" (p. 22). These skills, Blinder argues, include "creativity, inventiveness, spontaneity, flexibility, [and] interpersonal relations . . . not rote memorization" (p. 22). Similarly, in their book calling for a broader conceptualization of school accountability, Rothstein, Jacobsen, and Wilder (2008) highlight the broad scope of skills that students develop in school, including "the ability to reason and think critically, an appreciation of the arts and literature, . . . social skills and a good work ethic, good citizenship, and habits leading to good physical and emotional health."

This is not to say that value-added measurement cannot aid in evaluating certain basic – and even critically important – skills. Rather, they are simply too narrow to be relied upon as a meaningful representation of the range of skills, knowledge, and habits we expect teachers and schools to cultivate in their students.

Teachers or schools?

Even in cases where tests do adequately capture desired skills, it behooves us to ask whether value-added – a teacher's individual impact on students' academic progress – is, in fact, what is educationally relevant. Teachers certainly vary in effectiveness, and school leaders should be cognizant of their teachers' contribution to student success. Yet to the extent schooling is a group or team effort involving principals, teachers, and other school professionals (e.g., instructional coaches, librarians, counselors), Herculean efforts to isolate, report, and reward individual value-added ignores critical, interrelated parts of the educational process. At worst, narrow interest in individual results may undermine this process, a point I return to later.

This concern is hardly unique to education. Statistics on narrow metrics of individual productivity have their place in many organizations, from business and government to professional sports. Yet in most cases business leaders and athletic coaches recognize that the success of their organization is much more than the sum of their individual employee or player statistics (Rothstein 2009). HISD, and, to a lesser extent, the NYCDOE, with its small school-based performance bonus program (Springer & Winters 2009), have recognized that organizational outcomes are as important to recognize as individual successes. As value-added systems begin to be implemented in school systems nationwide, policy-makers should be aware of the potential educational costs of a narrow focus on individual metrics.

Is the measurement tool appropriate?

In assessing readily testable skills – fourth-grade arithmetic, for example – a measurement tool is needed that provides a valid and reliable assessment of students' mastery of the skill domain. No test will cover all skills that students are expected to master. By necessity, a test instrument must sample items from a much broader domain of skills. The resulting test may consist of, say, thirty-five to fifty multiple-choice questions. A student's performance on this test then provides an inference, approximation, or "educated guess" of his or her mastery of the broader skill domain (Koretz 2008).

A well-constructed test that draws evenly from the broader domain of skills is more likely to provide a valid inference about student learning. In the case of state tests – such as those administered in New York and Texas – the relevant domain is the state curriculum, which articulates what students should know and be able to do at each grade level. As noted above, however, many of these skills are not amenable to standardized testing and inevitably will be under-represented on the state test.[19]

Even for the standards that can be tested, many assessments are poor representations of these standards. Recent studies analyzing state test content in New York, Massachusetts, and Texas find that over many years of test administration, some parts of the state curriculum are never tested (Jennings & Bearak 2010; Holcolmbe, Jennings & Koretz 2010). To take one extreme case – the 2009 New York State eighth-grade math test – 50 percent of the possible points were based on only seven of the forty-eight state standards; what's more, only 51 percent of the points were required to pass.

Among skills that are consistently assessed, some receive predictably greater emphasis than others. Teachers aware of systematic omissions and repetitions can substantially inflate students' scores by narrowly focusing on frequently tested standards (popularly known as "teaching to the test").[20] For many tests, it is also possible to "teach to the format." Studies have illustrated how teachers focus their instruction on the format of the state test by presenting material in the same manner as it appears on the test (Darling-Hammond & Wise 1985; Shepard & Dougherty 1991; McNeil & Valenzuela 2000).[21] To the extent "teaching to the test" and "teaching to the format" behaviors differ across teachers – which they almost certainly do – true "value-added" comparisons will be compromised.

The fact that test items are sampled from a broader domain is relevant for more than just "gaming" behavior. Teachers vary in the extent to which their time and efforts align with content specifically emphasized on the state test,

[19] As two examples, the New York State standards for ELA include, "as speakers and writers, students will use oral and written language for self-expression and artistic creation" and "students will use oral and written language for effective communication with a wide variety of people." Few would disagree that these are important objectives; they are not, however, skills that are easily assessed on a standardized test.

[20] For example, when a novice math teacher's colleague informs her that the state test "never asks students to calculate the volume of a cylinder," that teacher can reliably improve scores by devoting less time on this concept.

[21] A striking example is reported in Shepard (1988). In that study, when a set of questions involving the addition of decimals was presented in a vertical format – as was standard on the state test – 86 percent of students answered these questions correctly. When the same problems were presented in a horizontal format, only 46 percent of students did.

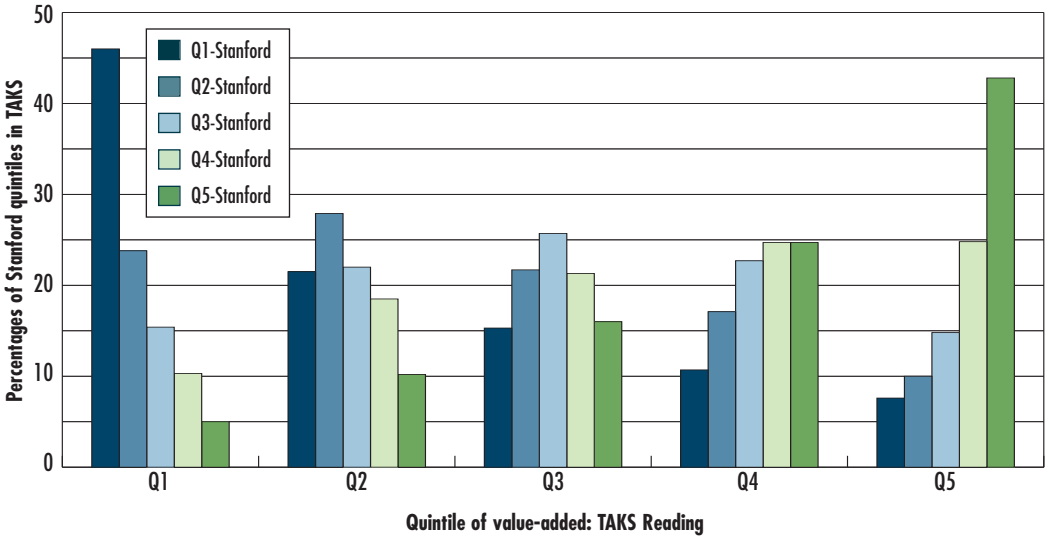
[22] These categories are quintiles based on value-added. Only students who took both the TAKS and Stanford Achievement Test are used to generate the value-added estimates.

for a variety of valid reasons. This variation may be due to the average ability level in their classroom, priorities of school leadership, parental demands, and so on. Given two teachers of equal effectiveness, the teacher whose classroom instruction happens to be most closely aligned with the test – for whatever reason – will outperform the other in terms of value-added.

Evidence that the choice of test can make a difference to value-added comes from recent research comparing value-added measures on multiple tests of the same content area. Since 1998, Houston has administered two standardized tests every year: the state TAKS and the nationally normed Stanford Achievement Test. Using HISD data, we calculated separate value-added measures for fourth- and fifth-grade teachers for the two tests (Corcoran, Jennings & Beveridge 2010). These measures were based on the same students, tested in the same subject, at approximately the same time of year, using two different tests.

We found that a teacher's value-added can vary considerably depending on which test is used. This is illustrated in Figure 5, which shows how teachers ranked on the two reading tests. Teachers are grouped into five performance categories on each test (1 to 5), with the five TAKS categories on the horizontal axis.[22] We see that teachers who had high value-added on one test tended to have high value-added on the other, but there were many inconsistencies. For example, among those who ranked in the top category (5) on the TAKS reading test, more than 17 percent ranked among the lowest two categories on the Stanford test. Similarly, more than 15 percent of the lowest value-added teachers on the TAKS were in the highest two categories on the Stanford.

Figure 5
Teacher value-added on two reading tests: Houston fourth- and fifth-grade teachers



Source: Corcoran, Jennings & Beveridge (2010)

These findings are consistent with those from other districts, including several large districts in Florida (Sass 2008) and a large anonymous district in the northeast (Papay 2010). To show what kind of an impact these inconsistencies might have in practice, Papay calculated hypothetical ASPIRE bonuses using his two differing sets of value-added estimates. He 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). He concludes that the two value-added estimates "are not sufficiently comparable to rank consistently the same individual teachers as high- or low-performing" (p. 3).

One of Papay's explanations for variation in value-added across tests is when the test was administered. That is, teachers' value-added looks different depending on whether the test was given in the early fall, mid-spring, or the end of the school year. Differences in test timing impact some teachers more than others, particularly those serving poor students who suffer a "summer learning loss" relative to their more advantaged peers (Alexander, Entwisle & Olsen 2001). When testing occurs in the middle of the year, value-added measures are made more difficult in that one has to apportion learning gains between two teachers. Until recently, New York State administered its grade three to eight exams in January and March; consequently, test score gains between two tests were due to two teachers, not one. It is not obvious how one appropriately isolates one teacher's impact from the other.

Can a teacher's unique effect be isolated?

As described in Section 2, the successful use of value-added in practice requires a high level of confidence in the attribution of achievement gains to specific teachers. One must be fairly confident that other explanations for test-score gains have been accounted for before rewarding or punishing teachers based on these measures. We saw earlier that something as simple as test timing can complicate the apportionment of gains between teachers. 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 teacher effects and classroom-level factors. In a given year, a class of students may perform particularly well or particularly poorly for reasons that have nothing to do with instruction. The proverbial "barking dog" on test day is one such explanation, as is a classroom illness or particularly disruptive student who affects the quality of instructional time. Over many years, this variation averages out, but in a single year the impact of the teacher cannot be separated from these influences. More years of test results helps, but this may be of little comfort to a teacher or school leader looking for actionable information today.

Most value-added models used in practice – including New York City's – also fail to separate teachers' influence from the school's effect on achievement.[23] That is, they don't account for the fact that performance differs systemically across schools due to differences in school policy, leadership, discipline, staff quality, and student mix. This omission is not simply an oversight. Value-added experts have pointed

out, rightly, that teacher effectiveness varies across schools within a district and to focus only on variation within schools would ignore important variation in teacher quality across schools (e.g., Gordon, Kane & Staiger 2006). The cost of this view, however, is that teacher effects end up confounded with school influences.

Recent research suggests that school-level factors can and do affect teachers' value-added. Jackson and Bruegmann (2009), for example, found in a study of North Carolina teachers that students perform better, on average, when their teachers have more effective colleagues. That is, Mrs. Appleton might have higher value-added when teaching next door to Mr. Johnson, because she benefits from his example, his mentoring, and his support. Other studies have found effects of principal leadership on student outcomes (Clark, Martorell & Rockoff 2009). Consequently, teachers rewarded or punished for their value-added may, in part, be rewarded or punished based on the teachers with whom they work.[24] This possibility certainly runs counter to the intended goal of value-added assessment.

Finally, as argued earlier, in many contexts, attempts to attribute achievement gains to individual teachers may not make sense in principle. This is most true in middle and high school, when students receive instruction from multiple teachers. To assume that none of these teachers' effects "spill over" into other coursework seems a strong – and unrealistic – assumption. Indeed, Koedel (2009) found that reading achievement in high school is influenced by both English and math teachers. Learning may simply not occur in the rigid way assumed by current value-added models.

Who counts?

Another significant limitation of value-added systems in practice is that they ignore a very large share of the educational enterprise. Not only do a minority of teachers teach tested subjects, but not all students are tested, and not all tested students contribute to value-added measures. In other words, from the standpoint of value-added assessment of teacher quality, these students do not count.[25]

In most states, including New York and Texas, students are tested in reading and mathematics annually in grades three to eight, and again in high school. Other subjects, including science and social studies, are tested much less often.[26] Because value-added requires a recent, prior measure of achievement in the same subject (usually last year's test score), only teachers of reading and math in grades four to eight can be assessed using value-added. Without annual tests, teachers cannot be assessed in other sub-

[23] Technically, the value-added model often does not include "school effects."

[24] In another study, Rothstein (2010) finds that a student's fifth-grade teacher has large effects on her fourth-grade achievement, a technical impossibility given that the student has not yet advanced to the fifth grade. He suggests that this finding may be due to "dynamic tracking," where a student's assignment to a fifth-grade teacher depends on their fourth-grade experience. When such assignment occurs, it biases measures of value-added.

[25] This is not a concern unique to teacher value-added measurement. The same issue arises when considering the design and effects of school accountability systems. When state testing systems (appropriately) allow exclusions for certain categories of students, incentives are created for schools to reclassify students such that they are exempted from the tests (see Figlio & Getzler 2002; Jacob 2005; and Jennings & Beveridge 2009).

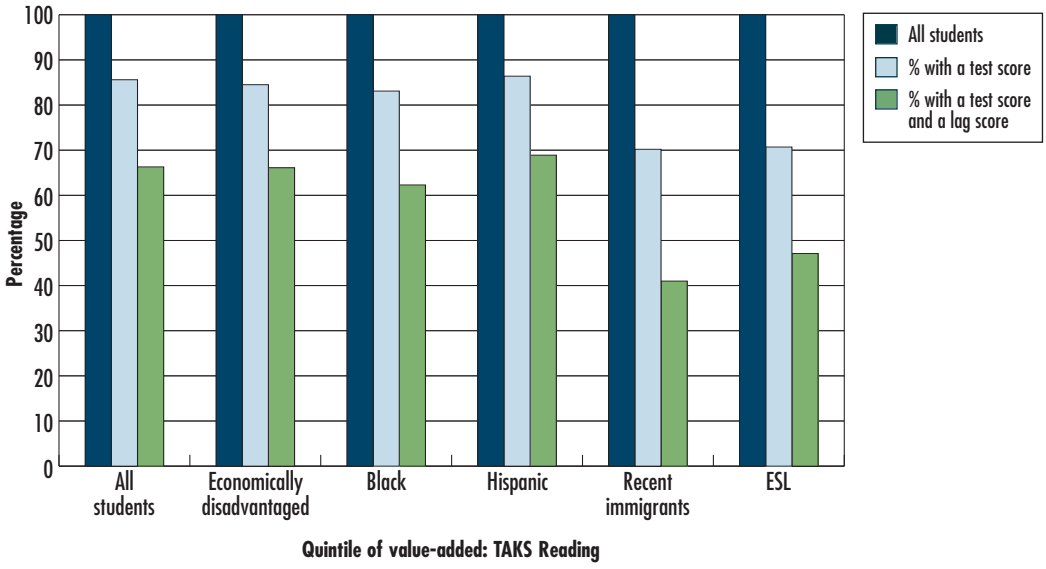
[26] In New York, students are tested in social studies in fifth and eighth grade, and science in fourth and eighth grade. See <www.emsc.nysed.gov/osa/schedules/2011/3-8schedule1011-021010.pdf>.

jects such as science and social studies. Thus, elementary, middle, and high school teachers of subjects other than reading and math are necessarily excluded from value-added assessment. Moreover, in many districts, students in grades six to eight attend middle schools where they receive instruction from multiple teachers. As discussed in the last section, attributing achievement gains to individual teachers in these settings is especially problematic.

Some students are routinely exempted from testing, or for one reason or another are missing a test score. Figure 6 illustrates how missing data can affect "who counts" toward a teacher's value-added assessment. This figure shows the percentage of students in grades four to six over eight years of testing in Houston who were tested in reading, and the percent of students who also had a test score for the prior year (and thus could contribute to a value-added estimate).[27] Due to disabilities, very low competency in the English language, absenteeism, and a variety of other reasons, approximately 14 percent of students in HISD did not have a test score in a typical year. This fraction varies by subgroup, depending on the extent of exemption in each group; for example, 15.6 percent of Black students failed to have a test score, and 26 to 29 percent of recent immigrants and ESL students were not tested.

Figure 6

Percent of students with a test score and percent contributing to value-added estimates, grades four to six, Houston,1998–2006



Source: Author's calculations using data from the Houston Independent School District. The percent of students with a test score and a lag score is only calculated for students in grades four to six (third grade is the first year of testing).

Because of high rates of student mobility in this population (in addition to test exemption and absenteeism), the percentage of students who have both a current and prior year test score – a prerequisite for value-added – is even lower (see Figure 6). Among all grade four to six students in HISD, only 66 percent had both of these scores, a fraction that falls to 62 percent for Black students, 47 percent for ESL students, and 41 percent for recent immigrants.

The issue of missing data is more than 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. Moreover, they provide little incentive for teachers to invest in students who will not count toward their value-added. Unfortunately, districts like New York City and Houston have very large numbers of highly mobile, routinely exempted, and frequently absent students. Moreover, 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.

[27] The latter is calculated only for students in grades four to six. Because third grade is the first year of testing, none of these students have a prior year score.

Are value-added scores precise enough to be useful?

As described in sections 2 and 3, value-added is based on a statistical model that effectively compares actual with predicted achievement. The residual gains serve as an estimate of the teacher's value-added. Like all statistical estimates, however, value-added has some level of uncertainty, or, a margin of error. In New York City's Teacher Data Reports, this uncertainty is expressed visually by a range of possible percentiles for each teacher's performance (the "confidence interval" for the value-added score). 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?

Let's return to the case of Mark Jones, whose data report is shown in Appendix B. Based on last year's test results, we learned that Mr. Jones ranked at the 43rd percentile among eighth-grade teachers in math. Taking into account uncertainty in this estimate, however, his range of plausible rankings range from the 15th to the 71st percentile. Although the 43rd percentile is our best estimate of Mr. Jones's performance, we can't formally rule out estimates ranging from 15 to 71. Using the NYCDOE performance categories, we can conclude that Mr. Jones is a "below average" teacher, an "average" teacher, or perhaps a borderline "above average" teacher.

What is the source of this uncertainty, exactly? 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 (and the more

narrow the estimate's confidence interval). With one year of data, it is impossible to separate teacher effects from other factors affecting the performance of students within a classroom (such as a disruptive classmate). While a particularly high set of test-score gains is suggestive of an effective teacher, one can only know if these gains are systematic after additional years of observations. These additional years are particularly important when scores are "noisy," that is, when achievement gains are not well explained by variables accounted for in the statistical model.[28]

Value-added estimates become more precise with additional years of data, as Mr. Jones's report illustrates. Based on his last four years of results, Mr. Jones ranked in the 56th percentile, though his plausible range still extends from the 32nd percentile to the 80th, which overlaps the "average" and "above average" performance categories. Taking his two sets of results together, we have learned that Mark Jones is most likely an average teacher, ranking somewhere around the 43rd and 56th percentile citywide, who may be below average, average, or above average. Accounting for uncertainty, we can't rule out a twenty-point margin in either direction.

It is unclear to this author what Mr. Jones or his principal can do with this information to improve instruction or raise student performance. More years of teaching experience allow the data to tell a slightly more precise story, but knowing that better data will be available in the future is of little use to Mr. Jones or his principal, who are looking for actionable information in real time. Value-added results for student subgroups would appear to be more promising to the extent they highlight subject areas or target populations in need of improvement – students who are English language learners, for example. Yet in most cases, the number of students used to calculate these subgroup estimates is so small that the resulting level of uncertainty renders them meaningless.

It is worth noting that – by design – 50 percent of teachers will perennially fall in the "average" performance category on the Teacher Data Report; another 40 percent will be considered "below average" or "above average. The remaining 10 percent are either "exceptional" (in the top 5 percent) or "failing" (in the bottom 5 percent). Thus, out of all teachers issued a value-added report in each year, half will be always be told little more than that they are "average."[29] At most, one in three will receive a signal to improve, though wide confidence intervals may and should raise doubt in the minds of some "below average" teachers. Of course, teachers who persistently score in the "high" category are probably doing something

[28] McCaffrey et al. (2009) show that much of the variation in teacher effects is due to random noise.

[29] It is often argued that policies of differentiating teacher effectiveness on the basis of test scores will lead to a long-run increase in the number of outstanding graduates willing to enter the teaching profession. Because highly effective teachers are not currently rewarded for their efforts – through higher salaries, for example, or promotion – candidates who are likely to be effective may be less likely to pursue a teaching career (Hoxby & Leigh 2004). However, there is currently little available evidence on either side of this question. It may turn out that the system for differentiating effectiveness matters. In the case of the NYCDOE Teacher Data Reports, only a modest fraction of teachers will be deemed "above average" or highly effective (top 5 percent). It could be that this has a discouraging effect on excellent teachers. In the words of a teacher who recently blogged about his data report results: "In fact, I'm mostly wholly average as an educator when it comes to teaching both math and reading. Not exactly the vote of confidence I was looking for. . . . This is disappointing to say the least. I did not join NYC Teaching Fellows to be an average teacher" (Brosbe 2010).

right and should be recognized; teachers persistently in the bottom 5 percent deserve immediate scrutiny. Still, it seems a great deal of effort has been expended to identify a small fraction of teachers. In the end, a tool designed for differentiating teacher effectiveness has done very little of the sort.

To get a better sense of the average level of uncertainty in the Teacher Data Reports, I examined the full set of value-added estimates reported to more than 12,700 teachers on the NYCDOE 2008-2009 reports. As we saw for Mark Jones in Appendix B, each value-added ranking is accompanied by a range of possible estimates. To begin, I simply calculated the width of this interval for every teacher in reading and math. Average widths across teachers are reported in Figure 7.

As expected, the level of uncertainty is higher when only one year of test results are used (the 2007-2008 bars) as against three years of data (all other bars). But in both cases, the average range of value-added estimates is very wide. For example, for all teachers of math, and using all years of available data, which provides the most precise measures possible, the average confidence interval width is about 34 points (i.e., from the 46th to 80th percentile). When looking at only one year of math results, the average width increases to 61 percentile points. That is to say, the average teacher had a range of value-added estimates that might extend from, for example, the 30th to the 91st percentile. The average level of uncertainty is higher still in ELA. For all teachers and years, the average confidence interval width is 44 points. With one year of data, this rises to 66 points.

Figure 7

Average confidence interval width, New York City Teacher Data Reports, 2008-2009

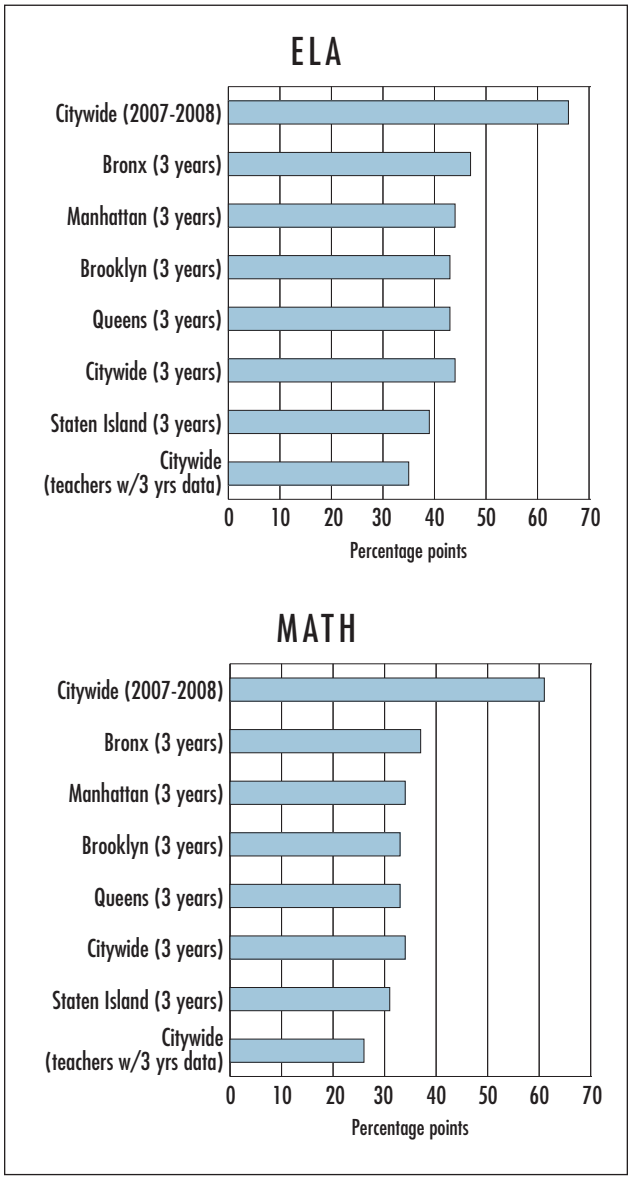
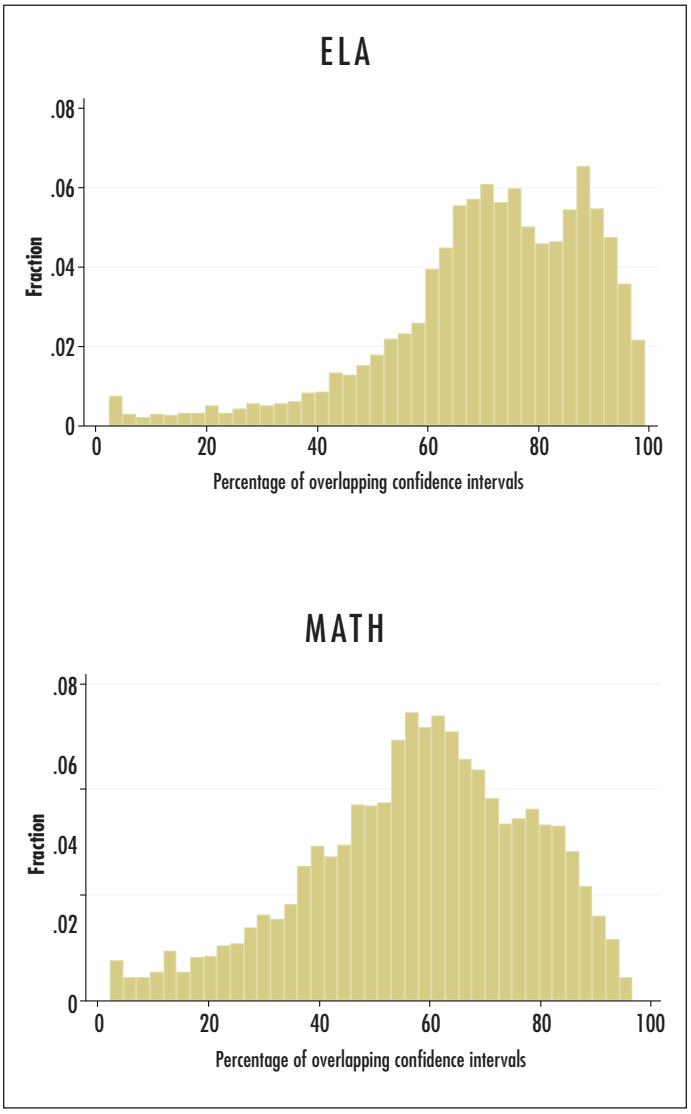


Figure 8
Percent of overlapping confidence intervals, ELA and math



Source: Author's calculations using data from the 2008-2009 Teacher Data Reports, based on up to three years of test results. Bars show the fraction of teachers that overlap with X percent of other teachers in the district teaching the same grade and subject.

As we saw earlier in this report, schools with high levels of mobility, test exemption, and absenteeism tend to have fewer students contributing to value-added estimates. And fewer student observations introduce a greater level of uncertainty associated with these estimates. Thus, a predictable pattern exists when comparing average levels of uncertainty across sections of the city. For example, the widest confidence intervals are found in the Bronx – whose schools serve many disadvantaged students – at 37 percentile points in math and 47 points in ELA (both based on up to three years of data; see Figure 7 on page 23). The most precise estimates, in contrast, are observed in relatively more advantaged Staten Island.

Another way of understanding the effects of uncertainty is to compare two teachers' ranges of value-added estimates and ask whether or not they overlap. For example, suppose that based on her value-added, Mrs. Appleton ranks in the 41st percentile of ELA teachers, with a confidence interval ranging from 24 to 58 (on the low end of the widths presented in Figure 7). And suppose Mr. Johnson ranks in the 51st percentile of ELA teachers, with an equally wide confidence interval from 34 to 68. Based on their "most likely" rankings, Mr. Johnson appears to have out-performed Mrs. Appleton. However, because we can't statistically rule out estimates in their overlapping intervals, we can't say with confidence that this is the case.

Using the 2008-2009 Teacher Data Report estimates, I compared all possible pairs of teacher percentile ranges in the city, within the same grade and subject, to see how many teachers could be statistically distinguished from one another.[30] For example, if Mrs.

Appleton's range of 24 to 58 overlaps with 56 percent of all other fourth-grade teachers, we could not rule out the possibility that she was equally effective as these 56 percent of teachers. The results are summarized in Figure 8, which shows the fraction of teachers who overlap with X percent of all other teachers. For example, the bar above 60 shows the fraction of teachers who cannot be statistically distinguished from 60 percent of all other teachers in the district.

Given the level of uncertainty reported in the data reports, half of teachers in grades three to eight who taught math have wide enough performance ranges that they cannot be statistically distinguished from 60 percent or more of all other teachers of math in the same grade. One in four teachers cannot be distinguished from 72 percent or more of all teachers. These comparisons are even starker for ELA, as seen in Figure 8. In this case, three out of four teachers cannot be statistically distinguished from 63 percent or more of all other teachers. Only a tiny proportion of teachers – about 5 percent in math and less than 3 percent in ELA – received precise enough percentile ranges to be distinguished from 20 percent or fewer other teachers.

As noted before, it is true that teachers' percentile ranking is their "best" or "most likely" estimate. But the ranges reported here cannot simply be ignored; they represent the extent of statistical precision with which the value-added estimate was calculated. Confidence intervals such as these are reported in any academic study that relies on inferential statistics, and any academic study that attempted to ignore these intervals in drawing between-group comparisons would in most cases be rejected outright.

[30] In other words, I compared every teacher's percentile range with every other teacher's percentile range. For example, if there are 1,000 teachers in the city, teacher 1's percentile range is compared to teachers 2 through 1,000; teacher 2's percentile range is compared with teachers 1 and 3 through 1,000, and so on, for all possible pairs.

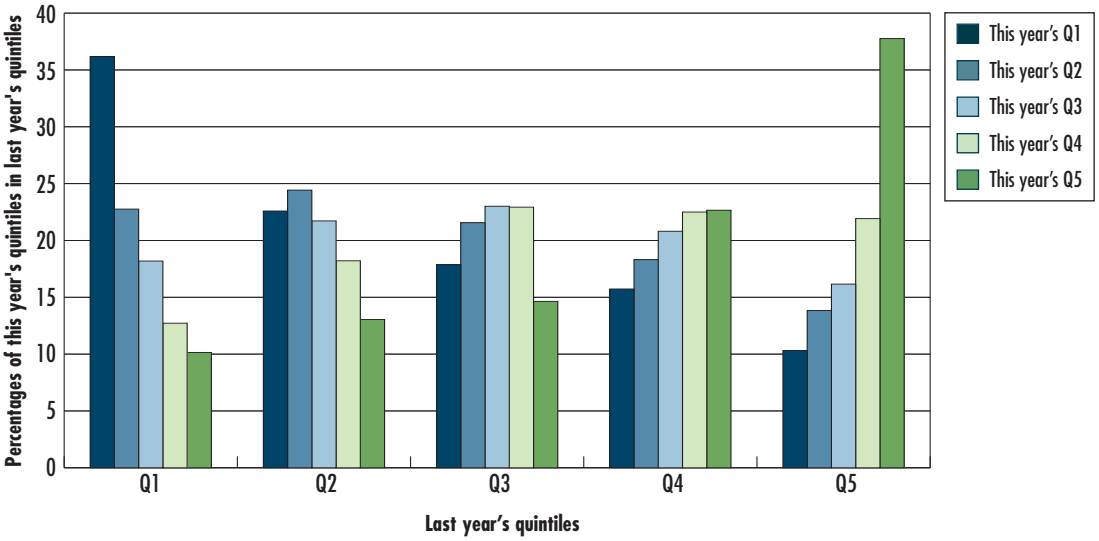
Is value-added stable from year to year?

Given the wide range of uncertainty observed in teacher value-added estimates, 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 the HISD data and the New York City Teacher Data Reports. Figure 9 shows how Houston teachers' value-added in reading correlates from one year to the next, using fourth- and fifth-grade data from 2000 to 2006. The format of this figure is very similar to that in Figure 5: Each bar shows the percent of teachers in each performance category in one year ("last year") that are in these same categories in the next year ("this year").

As in Figure 5, there is generally a positive correlation in teachers' value-added from one year to the next. For example, among those in the bottom quintile of performance last year, 36 percent remain in the bottom quintile in the following year. Similarly, among those in the top quintile of performance last year, 38 percent remain in the top quintile in the following year. Again, however, there are many inconsistencies. Twenty-three percent of last year's lowest performers are in the top two quintiles in the following year. Twenty-three percent of last year's highest performers are in the bottom two quintiles in the following year.

This variability from year to year would be expected of any statistical estimate that is estimated with error, and variability is reduced as additional years of data are added to the analysis. But here again, this may be of little comfort to a teacher who is using her annual estimates to improve her own practice. A top-performing teacher may be rewarded (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 the NYCDOE – caution users of this data against making rash decisions based on one year

Figure 9
Year-to-year stability in value-added rankings: HISD reading test, 2000–2006

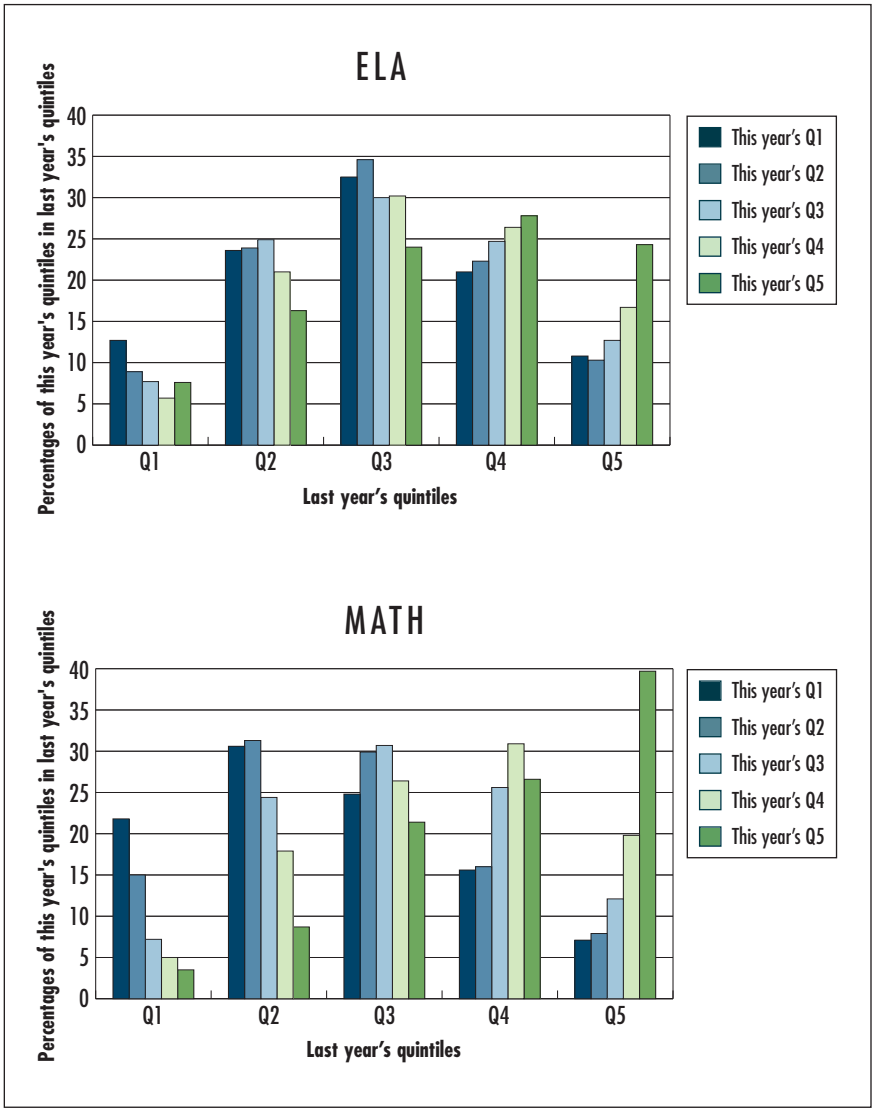


of estimates. But this estimate is one of only a few made available to teachers on their annual report, and thus they are hard to ignore. Inexperienced teachers – those arguably most in need of immediate feedback – simply will not have the multiple years of data on which to rely. It seems unlikely that teachers and their school leaders will not pay close attention to these noisy and imprecise estimates.

I repeated the analysis in Figure 9 for New York City's Teacher Data Reports, using teachers who had at least three years of value-added estimates, comparing value-added categories in 2007 and 2008 for both ELA and math. The results are presented in Figure 10. A pattern emerges that is nearly identical to that found for Houston, though the year-to-year correlation appears much weaker, especially in ELA. In math, about 23 percent of bottom quintile teachers in 2007 ranked in the top two quintiles in 2008. Top performers were much more consistent: 40 percent remained in the top quintile in 2008, while only 12 percent fell to the bottom two quintiles. ELA results were considerably more mixed. Most low performers in 2007 were not low performers in 2008, with some 31 percent ranking in the top two categories.

Figure 10

Year-to-year stability in ELA and math value-added rankings, New York City Teacher Data Reports, 2007-2008



5 Discussion

At least 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 their 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.

However, the promise that value-added systems can provide such a precise, meaningful, and comprehensive picture is not supported by the data. As the discussion in this report showed, 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 value-added assessment represents a small fraction of the goals our nation has set for our students and schools.

The implementation of value-added systems in practice 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, including New York City and Houston, 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. Recent 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, policymakers, and school leaders should not be seduced by the elegant simplicity of "value-added."

References

Alexander, Karl L., Doris R. Entwisle, and Linda S. Olsen. 2001. "Schools, Achievement, and Inequality: A Seasonal Perspective," *Educational Evaluation and Policy Analysis* 23:171–191.

Amrein-Beardsley, Audrey. 2008. "Methodological Concerns about the Education Value-Added Assessment System," *Educational Researcher* 37:65–75.

ASPIRE. 2010. "2008-2009 ASPIRE Award Program Highlights," <portal.battelleforkids.org/ASPIRE/Recognize/ASPIRE_Award/2009_aspire_highlights.html>.

Associated Press. 2007. "School District Asks Teachers to Return Pay," *New York Times* (March 11).

Blinder, Alan S. 2009. "Education for the Third Industrial Revolution." In *Creating a New Teaching Profession*, edited by Dan Goldhaber and Jane Hannaway, pp. 3–14. Washington, DC: Urban Institute Press.

Blumenthal, Ralph. 2006. "Houston Ties Teacher Pay to Student Test Scores," *New York Times* (January 13).

Boyd, Donald, Pamela Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff. 2006. "How Changes in Entry Requirements Alter the Teacher Workforce and Affect Student Achievement," *Education Finance and Policy* 1:176–216.

Boyd, Donald J., Pamela L. Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff. 2009. "Teacher Preparation and Student Achievement," *Educational Evaluation and Policy Analysis* 31:416–440.

Boyd, Donald J., Hamilton Lankford, Susanna Loeb, Jonah Rockoff, and James Wyckoff. 2008. "The Narrowing Gap in New York City Teacher Qualifications and Its Implications for Student Achievement in High-Poverty Schools," *Journal of Policy Analysis and Management* 27:793–818.

Braun, Henry I. 2005. *Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models.* Policy Information Perspective. Princeton, NJ: Educational Testing Service.

Brosbe, Ruben. 2010. "My Disappointing Data and What to Do With It," Gotham Schools Classroom Tales Blog (March 10), <gothamschools.org/2010/03/10/my-disappointing-data-and-what-to-do-with-it/>.

Buddin, Richard, Daniel F. McCaffrey, Sheila Nataraj Kirby, and Nailing Xia. 2007. "Merit Pay for Florida Teachers: Design and Implementation Issues." RAND Education Working Paper, WR-508-FEA. Santa Monica, CA: RAND.

Center for Educator Compensation Reform. 2008. *Performance Pay in Houston.* Washington DC: U.S. Department of Education. Downloadable PDF at <www.cecr.ed.gov/guides/summaries/HoustonCaseSummary.pdf>.

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

Clotfelter, Charles T., Helen F. Ladd, and Jacob Vigdor. 2005. "Who Teaches Whom? Race and the Distribution of Novice Teachers," *Economics of Education Review* 24:377–392.

Corcoran, Sean P., Jennifer L. Jennings, and Andrew A. Beveridge. 2010. "Teacher Effectiveness on High- and Low-Stakes Tests." Paper presented at the Institute for Research on Poverty summer workshop, Madison, WI.

Darling-Hammond, Linda, and Alfred E. Wise. 1985. "Beyond Standardization: State Standards and School Improvement," *Elementary School Journal* 85:315–336.

Figlio, David N., and Lawrence S. Getzler. 2002. "Accountability, Ability and Disability: Gaming the System." Working Paper No. 9307. Cambridge, MA: National Bureau of Economic Research.

Goldhaber, Dan. 2008. "Teachers Matter, but Effective Teacher Policies are Elusive." In *Handbook of Research in Education Finance and Policy*, edited by Helen Ladd, and Edward B. Fiske. New York: Routledge.

Gordon, Robert, Thomas J. Kane, and Douglas O. Staiger. 2006. *Identifying Effective Teachers Using Performance on the Job.* The Hamilton Project White Paper 2006-01. Washington, DC: The Brookings Institution.

Hanushek, Eric A., and Steven G. Rivkin. 2006. "Teacher Quality." In *Handbook of the Economics of Education*, edited by E. A. Hanushek and F. Welch. Amsterdam, The Netherlands: Elsevier.

Harris, Douglas N. 2009. "Teacher Value-Added: Don't End the Search Before it Starts," *Journal of Policy Analysis and Management* 28:693–699.

Hill, Heather C. 2009. "Evaluating Value-Added Models: A Validity Argument Approach," *Journal of Policy Analysis and Management* 28:700–709.

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.

Hoxby, Caroline M., and Andrew Leigh. 2004. "Pulled Away or Pushed Out? Explaining the Decline of Teacher Aptitude in the United States," *American Economic Review* 94:236–240.

Houston Independent School District, Department of Research and Accountability. 2009. *2005-2006 Teacher Performance Pay and 2006-07 ASPIRE Award Program Evaluation.* Houston, TX: HISD.

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.

Jacob, Brian A. 2005. "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools," *Journal of Public Economics* 89:761–796.

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

Jennings, Jennifer L., and Andrew A. Beveridge. 2009. "How Does Test Exemption Affect Schools' and Students' Academic Performance?" *Educational Evaluation and Policy Analysis* 31:153–175.

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.

Keller, Bess. 2008. "Drive On to Improve Evaluation Systems for Teachers," *Education Week* (January 15).

Klein, Joel, and Weingarten, Randi. 2008. Joint letter. Principals' Weekly (October 1).

Koedel, Cory. 2009. "An empirical analysis of teacher spillover effects in secondary school," *Economics of Education Review* 28:682–692.

Koretz, Daniel. 2008. "A Measured Approach," *American Educator* (Fall), 18–27, 39.

Lankford, Hamilton, Susanna Loeb, and James Wyckoff. 2002. "Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis," *Educational Evaluation and Policy Analysis* 24:37–62.

Martinez, Barbara. 2010. "School Tenure Crackdown," *Wall Street Journal* (July 30).

McCaffrey, Daniel F., Daniel M. Koretz, J. R. Lockwood, and Laura S. Hamilton. 2004. *Evaluating Value-Added Models for Teacher Accountability.* Santa Monica, CA: RAND.

McNeil, Linda, and Angela Valenzuela. 2000. *The Harmful Impact of the TAAS System of Testing in Texas: Beneath the Accountability Rhetoric.* Educational Resources Information Center Report ED 443-872. Washington, DC: U.S. Department of Education.

McCaffrey, Daniel F., Tim R. Sass, J. R. Lockwood, and Kata Mihaly. 2009. "The Intertemporal Variability of Teacher Effect Estimates," *Education Finance and Policy* 4:572–606.

Medina, Jennifer. 2008a. "Bill Would Bar Linking Class Test Scores to Tenure," *New York Times* (March 18).

Medina, Jennifer. 2008b. "New York Measuring Teachers by Test Scores," *New York Times* (January 21).

Medina, Jennifer. 2009. "A Two-Year Study to Learn What Makes Teachers Good," *New York Times City Room Blog* (September 1), <http://cityroom.blogs.nytimes.com/2009/09/01/a-2-year-study-to-learn-what-makes-teachers-good/>.

Medina, Jennifer. 2010. "Agreement Will Alter Teacher Evaluations," *New York Times* (May 10).

Mellon, Ericka. 2010a. "HISD to Pay Out More Than $40 Million in Bonuses," *Houston Chronicle* (January 27).

Mellon, Ericka, 2010b. "HISD Spells Out Teacher Dismissal Process, Part II," *Houston Chronicle School Zone Blog* (February 8), <http://blogs.chron.com/schoolzone/2010/02/hisd_spells_out_teacher_dismis.html>.

Murnane, Richard J., and David K. Cohen. 1986. "Merit Pay and the Evaluation Problem: Why Most Merit Pay Plans Fail and a Few Survive," *Harvard Educational Review* 56:1–17.

Nye, Barbara, Spyros Konstantopoulos, and Larry V. Hedges. 2004. "How Large Are Teacher Effects?" *Educational Evaluation and Policy Analysis* 26:237–257.

Obama, Barack. 2009. "Remarks by the President on Education (July 24)." Press Release. Washington DC: Office of the Press Secretary. Available online at: <www.whitehouse.gov/the_press_office/remarks-by-the-president-at-the-department-of-education>

Odden, Allan, and Carolyn Kelley. 2002. *Paying Teachers for What They Know and Do: New and Smarter Compensation Strategies to Improve Schools*, 2nd ed. Thousand Oaks, CA: Corwin Press.

Olson, Lynn. 2007. "Houston Overhauls Merit Pay System," *Education Week* (September 14).

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.*

Rivkin, Steven G. 2007. *Value-Added Analysis and Education Policy.* CALDER Policy Brief No. 1. Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. "Teachers, Schools, and Academic Achievement," *Econometrica* 73:417–458.

Rockoff, Jonah E. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," American Economic Review 94:247–252.

Rockoff, Jonah E. 2008. "Does Mentoring Reduce Turnover and Improve Skills of New Employees? Evidence from Teachers in New York City." Working Paper No. 16240. Cambridge: National Bureau of Economic Research.

Rockoff, Jonah E., Douglas O. Staiger, Thomas J. Kane, and Eric S. Taylor. 2010. "Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools." Working Paper. New York: Columbia Business School.

Rothstein, Jesse. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics* 125:175–214.

Rothstein, Richard. 2009. "The Influence of Scholarship and Experience in Other Fields on Teacher Compensation Reform." In *Performance Incentives: Their Growing Impact on American K-12 Education*, edited by Matthew G. Springer. Washington, DC: The Brookings Institution.

Rothstein, Richard, Rebecca Jacobsen, and Tamara Wilder. 2008. *Grading Education: Getting Accountability Right.* Washington, DC and New York: Economic Policy Institute and Teachers College Press.

Sanders, William L., and Sandra P. Horn. 1994. "The Tennessee Value-Added Assessment System (TVAAS): Mixed-Model Methodology in Educational Assessment," *Journal of Personnel Evaluation in Education* 8:299–311.

Sanders, William L., Arnold M. Saxton, and Sandra P. Horn. 1997. "The Tennessee Value-Added Assessment System: A Quantitative-Based Approach to Educational Assess-

ment." In *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure?* edited by Jason Millman. Thousand Oaks, CA: Corwin Press, Inc.

Sawchuk, Stephen. 2010. "Houston Approves Use of Test Scores in Teacher Dismissals," *Education Week Teacher Beat Blog* (February 12), <http://blogs.edweek.org/edweek/teacherbeat/2010/02/houston_approves_use_of_test_s.html>.

Sass, Tim R. 2008. *The Stability of Value-Added Measures of Teacher Quality and Implications for Teacher Compensation Policy.* CALDER Policy Brief #4. Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Shepard, Lorrie A. 1988. "The Harm of Measurement-Driven Instruction," Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC.

Shepard, Lorrie A., and Katherine Dougherty. 1991. "Effects of High-Stakes Testing on Instruction." Paper presented at the annual meetings of the American Educational Research Association and the National Council of Measurement in Education, Chicago, IL.

Springer, Matthew G., and Marcus A. Winters. 2009. *The NYC Teacher Pay-for-Performance Program: Early Evidence from a Randomized Trial.* Civic Report No. 56. New York: Manhattan Institute.

Toch, Thomas, and Robert Rothman. 2008. *Rush to Judgment: Teacher Evaluation in Public Education.* Washington, DC: Education Sector.

U.S. Department of Education. 2010. "Overview Information: Race to the Top Fund; Notice Inviting Applications for New Awards for Fiscal Year (FY) 2010," *Federal Register* 75, no. 71 (April 14), Part III: U.S. Department of Education, pp. 19,499–19,500. Washington DC: U.S. GPO. Downloadable PDF at: <www2.ed.gov/legislation/FedRegister/announcements/2010-2/041410a.pdf>.

Weisburg, Daniel, Susan Sexton, Susan Mulhern, and David Keeling. 2009. *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness.* Brooklyn, NY: New Teacher Project.

APPENDIX A

Race to the Top Definitions of Teacher Effectiveness and Student Achievement

Effective teacher means a teacher whose students achieve acceptable rates (e.g., at least one grade level in an academic year) of student growth (as defined in this notice). States, LEAs, or schools must include multiple measures, provided that teacher effectiveness is evaluated, in significant part, by student growth (as defined in this notice). Supplemental measures may include, for example, multiple observation-based assessments of teacher performance.

Highly effective teacher means a teacher whose students achieve high rates (e.g., one and one-half grade levels in an academic year) of student growth (as defined in this notice). States, LEAs, or schools must include multiple measures, provided that teacher effectiveness is evaluated, in significant part, by student growth (as defined in this notice). Supplemental measures may include, for example, multiple observation-based assessments of teacher performance or evidence of leadership roles (which may include mentoring or leading professional learning communities) that increase the effectiveness of other teachers in the school or LEA.

Student achievement means (a) For tested grades and subjects: (1) A student's score on the State's assessments under the ESEA; and, as appropriate, (2) other measures of student learning, such as those described in paragraph (b) of this definition, provided they are rigorous and comparable across classrooms. (b) For non-tested grades and subjects: Alternative measures of student learning and performance such as student scores on pre-tests and end-of-course tests; student performance on English language proficiency assessments; and other measures of student achievement that are rigorous and comparable across classrooms.

Student growth means the change in student achievement (as defined in this notice) for an individual student between two or more points in time. A State may also include other measures that are rigorous and comparable across classrooms.

Source: U.S. Department of Education, "Overview Information: Race to the Top Fund; Notice Inviting Applications for New Awards for Fiscal Year (FY) 2010," *Federal Register* 75, no. 71 (April 14, 2010), Part III: U.S. Department of Education, pp. 19,499–19,500, Washington DC: U.S. GPO. Downloadable PDF at: <www2.ed.gov/legislation/FedRegister/announcements/2010-2/041410a.pdf>.

Sample New York City Teacher Data Report, 2010

I.S. 000
8th Grade
Years Teaching in NYC: 4+ years

NYC Department of Education TEACHER DATA REPORT

MATH MARK JONES

Teacher Data Reports–a different perspective

Every day, you and your principal see the impact you have on your students' learning in multiple ways, including student work, student feedback, and classroom observation.

Sometimes, though, it's helpful to understand how your impact compares with teachers across the city. Teacher Data Reports provide that different perspective.

How to interpret the data: SAMPLE

	NUMBER OF STUDENTS	PRIOR PROFICIENCY	PREDICTED SCORE	ACTUAL SCORE	VALUE ADDED	PERCENTILE (RANGE
2008-09	34	3.10	3.33	3.52	0.19	82

The teacher shown in the chart above had **34 students** in her class contribute to her **value-added** score in 2008-2009. The average **prior proficiency rating** of these students was a **3.10**. Across the city, students that are similar to these students scored a 3.33 on the next test, so the predicted score for these students is 3.33. These students **scored a 3.52**, on average, on the next test. Thus, this teacher was able to move these students more than other teachers were able to move similar students by 0.19 of a proficiency level. This number is the teacher's value-added. A **value-added of 0.19** puts this teacher in the **82nd percentile**, which means her value score is higher than 82% of the teachers in the comparison group. Below the teacher's percentile is a line representing the range — we can be 95% certain that this teacher's result is somewhere on this line, most likely towards the center.

How to interpret performance categories

PERFORMANCE CATEGORY	
HIGH	**Above** the 95th percentile
ABOVE AVG	Between the 75th-95th percentile
AVERAGE	Between the 25th-75th percentile
BELOW AVG	Between the 5th-25th percentile
LOW	**Below** the 5th percentile

WHAT DATA IS BEING USED?
- Standardized NYS tests in math and ELA
- The number of years you've taught in NYC schools
- Student and classroom characteristics that can be related to student achievement gains

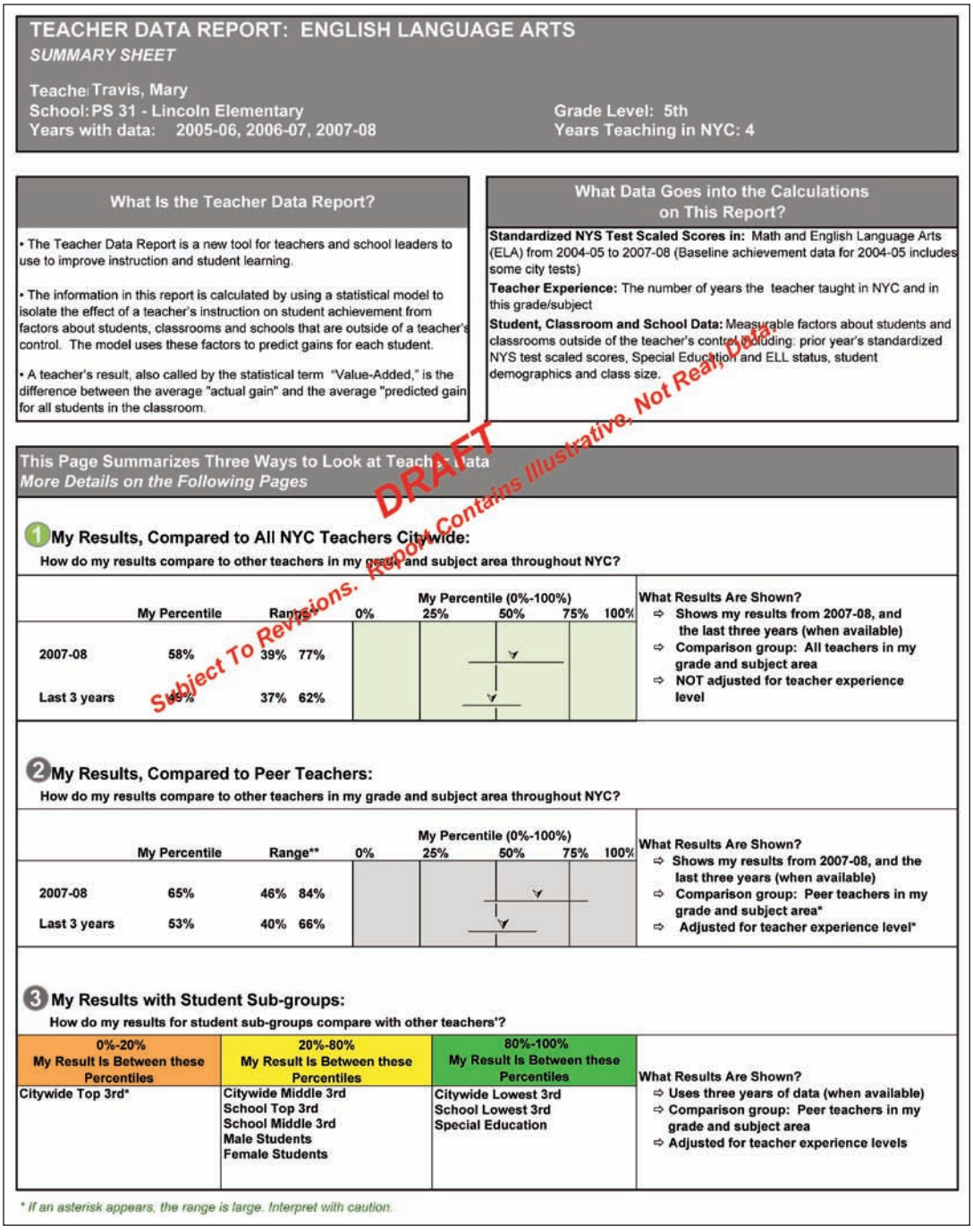
How do my results compare to other teachers?

The chart below shows how your results compare to other teachers who teach the same grade, the same subject area and have a similar amount of experience.

	NUMBER OF STUDENTS	PRIOR PROFICIENCY	PREDICTED SCORE	ACTUAL SCORE	VALUE ADDED	MY PERCENTILE / (RANGE) 0% 25% 50% 75% 100%	PERFORMANCE CATEGORY
My Results							
2008-2009	27	3.27	3.29	3.26	-0.02	43	AVERAGE
Last 4 Years	61	3.00	2.92	2.95	0.03	56	AVERAGE

Source: New York City Department of Education

APPENDIX C Sample New York City Teacher Data Report, 2009



TEACHER DATA REPORT: ENGLISH LANGUAGE ARTS

SUMMARY SHEET

Teacher: Travis, Mary
School: PS 31 - Lincoln Elementary
Years with data: 2005-06, 2006-07, 2007-08
Grade Level: 5th
Years Teaching in NYC: 4

What Is the Teacher Data Report?

- The Teacher Data Report is a new tool for teachers and school leaders to use to improve instruction and student learning.
- The information in this report is calculated by using a statistical model to isolate the effect of a teacher's instruction on student achievement from factors about students, classrooms and schools that are outside of a teacher's control. The model uses these factors to predict gains for each student.
- A teacher's result, also called by the statistical term "Value-Added," is the difference between the average "actual gain" and the average "predicted gain for all students in the classroom.

What Data Goes into the Calculations on This Report?

Standardized NYS Test Scaled Scores in: Math and English Language Arts (ELA) from 2004-05 to 2007-08 (Baseline achievement data for 2004-05 includes some city tests)

Teacher Experience: The number of years the teacher taught in NYC and in this grade/subject

Student, Classroom and School Data: Measurable factors about students and classrooms outside of the teacher's control including: prior year's standardized NYS test scaled scores, Special Education and ELL status, student demographics and class size.

DRAFT
Subject To Revisions. Report Contains Illustrative, Not Real, Data.

This Page Summarizes Three Ways to Look at Teacher Data
More Details on the Following Pages

1 My Results, Compared to All NYC Teachers Citywide:

How do my results compare to other teachers in my grade and subject area throughout NYC?

	My Percentile	Range*
2007-08	58%	39% 77%
Last 3 years	49%	37% 62%

My Percentile (0%-100%): 0% 25% 50% 75% 100%

What Results Are Shown?
- Shows my results from 2007-08, and the last three years (when available)
- Comparison group: All teachers in my grade and subject area
- NOT adjusted for teacher experience level

2 My Results, Compared to Peer Teachers:

How do my results compare to other teachers in my grade and subject area throughout NYC?

	My Percentile	Range**
2007-08	65%	46% 84%
Last 3 years	53%	40% 66%

My Percentile (0%-100%): 0% 25% 50% 75% 100%

What Results Are Shown?
- Shows my results from 2007-08, and the last three years (when available)
- Comparison group: Peer teachers in my grade and subject area*
- Adjusted for teacher experience level*

3 My Results with Student Sub-groups:

How do my results for student sub-groups compare with other teachers'?

0%-20% My Result Is Between these Percentiles	20%-80% My Result Is Between these Percentiles	80%-100% My Result Is Between these Percentiles
Citywide Top 3rd*	Citywide Middle 3rd School Top 3rd School Middle 3rd Male Students Female Students	Citywide Lowest 3rd School Lowest 3rd Special Education

What Results Are Shown?
- Uses three years of data (when available)
- Comparison group: Peer teachers in my grade and subject area
- Adjusted for teacher experience levels

** If an asterisk appears, the range is large. Interpret with caution.*

Source: New York City Department of Education

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