

Can NYC Teachers be Evaluated by Student Test Scores? *Should They Be?*

Sean P. Corcoran

New York University
Institute for Education and Social Policy
Steinhardt School of Culture, Education and Human Development &
Wagner School of Public Service

Annenberg Institute for School Reform - NYC Conversation Series
January 27, 2010

Introduction

- ❖ Teachers are likely the most important *school* influence on student achievement
- ❖ There is strong evidence that teachers **vary widely in effectiveness**, and that teacher quality is inequitably distributed across students and schools
- ❖ There is little disagreement that **teacher quality is critical** to students' and schools' success

Introduction

- ❖ **How should we measure teacher quality, in practice?**
- ❖ **How should we address, develop, and support under-performing teachers?**
- ❖ There are a number of ways to evaluate teachers' performance, but most in use rely on (infrequent) classroom observations

Introduction

- ❖ **Value-added methods** that use test score gains to try to isolate teachers' unique contribution to student learning are seen as a promising tool for identifying teaching effectiveness

"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." *President Barack Obama, July 24, 2009*

Introduction: “Race to the Top”

- ❖ ≥ 72 of 500 points devoted to the use of student achievement to assess teacher quality
- ❖ FL, IN, RI, TN: at least 51% of teacher and principal evaluations tied to student test scores
- ❖ Houston Independent School District has approved a policy to dismiss teachers in part based on value-added scores

My goals for this presentation

1. To introduce value-added methods: *what is a teacher’s “value-added?”*
2. To introduce NYC’s application of value-added: the Teacher Data Initiative
3. To provide an overview of the challenges facing this particular form of teacher evaluation

My goals for this presentation

4. To begin a discussion around the evaluation of teacher effectiveness in New York City:
 - Can we evaluate individual teachers using test scores?
 - Should we?
 - What are the alternatives?
 - Is value-added “better than the status quo?”

What is a teacher’s “value-added?”

- ❖ Suppose we wanted to use standardized test results to compare two 4th grade teachers:
 - Mrs. Appleton’s class: avg. math score of 42
 - Mr. Johnson’s class: avg. math score of 75
 - *Who is more effective?*
- ❖ Average test score *levels* are a reflection of student background, school and out-of-school factors

What is a teacher's "value-added?"

- ❖ A better approach might be to look at their students' *progress* in math from the 3rd grade:
 - Ms. Appleton's students improved **10 pts on average**
 - Mr. Johnson's students improved **4 pts on average**
 - *Who is more effective?*
- ❖ Test score *gains* get us closer to a teacher "effect," but what part of this gain can be attributed solely to the teacher?

What is a teacher's "value-added?"

- ❖ If students were **randomly assigned** to teachers, this would be easy: systematic differences would almost surely be due to teacher assignment
 - Of course, they aren't
- ❖ So, the goal becomes devising a statistical model **to account for other factors** that explain differences in achievement

What is a teacher's "value-added?"

A teacher's "value-added" is her students' **average test score gain**, *properly adjusted*

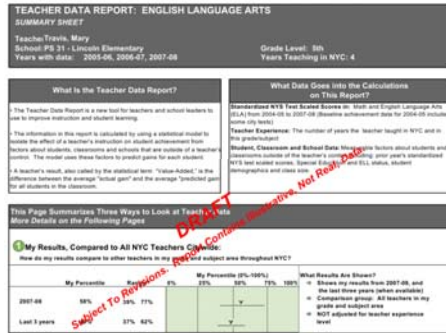
- ❖ To be useful for practice, we need a high level of confidence in attribution:
 - *How would Ms. Appleton's students have fared if they had not had her as a teacher?*
 - It takes a lot of information and assumptions to be confident of the answer to this question

New York City

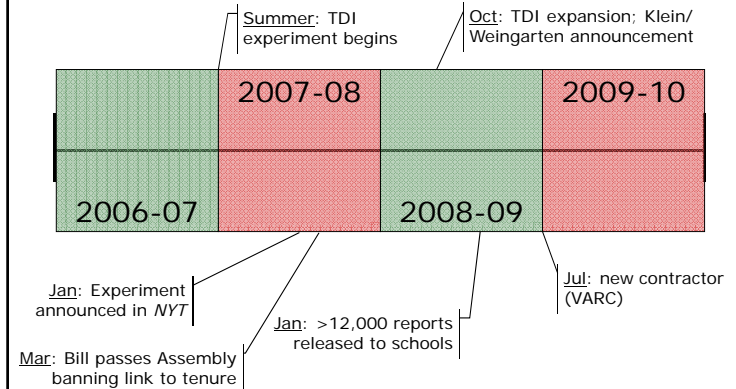
- ❖ New York City has been at the forefront of the development and use of value-added measurement for research and practice
 - Student test scores linked to teachers
 - Research on characteristics associated with teacher effectiveness (e.g. certification, experience, TFA)
 - **Teacher Data Initiative (TDI)**

Teacher Data Initiative

- ❖ One value-added report per subject for all teachers of **ELA and mathematics, grades 4-8**



Teacher Data Initiative



Teacher Data Initiative

“The [NYC] mayor...announced that he has instructed City Schools Chancellor Joel I. Klein to begin **using student performance data immediately** to inform teacher tenure decisions” *Press release, NYC Office of the Mayor, November 25, 2009*

Teacher Data Initiative: DoE goals

- ❖ To offer “**one lens on teacher effectiveness**” to be triangulated with other information about teacher quality (can serve as a “red flag”)
- ❖ To **stimulate conversation** about student achievement and promote better instructional practices through professional development
- ❖ To learn more about “what works” in the classroom

Teacher Data Initiative: DoE goals

- ❖ Teacher Data Reports are an *evolving tool*
 - New statistical methodology for 2009-10
 - Simplified presentation
 - Responsive to feedback from principals

- ❖ New York City is taking part in a large-scale study based at Harvard on multiple measures of teacher effectiveness
 - Measuring Effective Teaching (metproject.org)

What a teacher data report includes

- ❖ Value-added measures (percentiles) based on:
 - Last year's test score gains
 - Multiple years of test score gains (where available)

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**	My Percentile (0%-100%)				What Results Are Shown?
			0%	25%	50%	75%	
2007-08	65%	46% 84%			v		☐ 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*
Last 3 years	53%	40% 66%			v		

What a teacher data report includes

- ❖ And relative to:
 - All teachers citywide
 - A comparison group of "peer" teachers

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**	My Percentile (0%-100%)				What Results Are Shown?
			0%	25%	50%	75%	
2007-08	58%	39% 77%			v		☐ 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
Last 3 years	37%	22% 62%			v		

*Subject To Revisions. Report CO**

What a teacher data report includes

- ❖ And for *subgroups* of students:
 - ELL, special ed, gender, initial achievement; not race

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	What Results Are Shown?
Citywide Top 3rd*	Citywide Middle 3rd	School Top 3rd School Middle 3rd Male Students Female Students	Citywide Lowest 3rd School Lowest 3rd Special Education	☐ 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.

How it works

- ❖ Suppose Mrs. Appleton has **25** 4th grade students
- ❖ A statistical model produces a *predicted test score* in ELA (and math) for each student in Mrs. Appleton's class, based on a number of things...

How it works

Student characteristics	Classroom characteristics	School characteristics
<ul style="list-style-type: none"> ✓ Prior year reading ✓ Prior year math ✓ Free or reduced price lunch ✓ Special education status ✓ English Learner status ✓ Number of suspensions and absences (prior-year) ✓ Student retained in grade ✓ Attended summer school ✓ New to school ✓ Race ✓ Gender ✓ Prior year teacher 	<ul style="list-style-type: none"> ✓ Average prior year reading and math ✓ Percent free or reduced price lunch ✓ Percent special education status ✓ Percent English Learner status ✓ Average number of suspensions and absences (prior) ✓ Percent of students retained in grade ✓ Percent attended summer school ✓ Class size ✓ Percent by race ✓ Percent by gender 	<ul style="list-style-type: none"> ✓ Average classroom characteristics ✓ Average class size ✓ Total tested by grade/subject ✓ Year starting and ending school <p><u>Teacher Characteristics</u> (used when comparing teachers to peer teachers)</p> <ul style="list-style-type: none"> ✓ Years of experience ✓ Years teaching in the same grade and subject

Source: NYC DoE training presentation, May 2008

How it works

- ❖ The extent to which Mrs. Appleton's students systematically do *better or worse than predicted* is a measure of her *value-added*
- ❖ That is, it's her "students' average test score gain, *properly adjusted*"

How it works

- ❖ There will be a *distribution* of value-added; some classrooms will do much better than predicted, and others worse
- ❖ Each teacher's *percentile rank* among all teachers, or among her peer group is reported

How it works

2 My Results, Compared to Peer Teachers:

How do my results compare to other teachers in my grade and subject area throughout NYC, whose classrooms have similar predicted gains, adjusted for teacher experience levels?

- ↳ 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

	Number of Students	Prior Proficiency Rating	Average			Percentile (0-100%)	My Percentile				
			Actual Gain	Predicted Gain	Value-Added		0%	25%	50%	75%	100%
This year: 2007-08	24	2.1	0.19	0.08	0.11	65%	▼				
Range						46-84%					
2006-07	24	2.4	0.08	0.08	0.00	50%	▼				
Range						30-70%					
2005-06	25	2.5	0.03	0.06	-0.03	43%	▼				
Range						22-64%					
Last 3 years average	73	2.4	0.10	0.07	0.03	53%	▼				
Range						40-66%					

Value-added: challenges

- ❖ Value-added measurement is promising and **intuitively appealing**, but susceptible to misuse
 - Has been challenged on a number of grounds
- ❖ Before adopting value-added measurement in practice, we need to be aware of its limitations— and “ask the hard questions”
 - Fortunately, this is a very active area of research

Value-added: challenges

1. **What** are we measuring?
2. Is **the measurement tool** appropriate?
3. Can we really isolate teachers' **unique contribution**?
4. **Who counts**?
5. Are estimates **precise enough** to be meaningful?
6. Are teacher effects **stable** from year to year?
7. Do rewards and sanctions based on value-added create appropriate **incentives**?

(1) What are we measuring?

- ❖ What do we expect students to know and be able to do?
- ❖ Standardized tests represent **only a subset of educational goals**, and even then focus on achievement in a few subject areas
- ❖ Within subjects, only a subset of skills are typically tested (or *can* easily be tested)

(1) What are we measuring?

- ❖ How do schools promote desired outcomes?
 - A **collective effort**, or an individual responsibility?
 - What do we expect **teachers** to do?
- ❖ What other **school or district** inputs matter?
- ❖ Does value-added shift the focus away from organizational responsibility?

(2) The measurement tool

- ❖ Which skills are represented on the test?
 - A test is often 35-50 questions
 - Classrooms, schools may—purposefully—have different emphases
 - E.g. 2009 NYS 8th grade math test—50 percent of the test points were taken from 7 standards (of 48) and only 51% needed to pass

(2) The measurement tool

- ❖ When is the test given? Test timing has large implications for value-added
 - NY administered its grade 3-8 exams in January and March (now April and May)
- ❖ Does the test design matter?
 - Research finds value-added measures can differ substantially for the same teacher across different tests of the same subject

(3) Isolating teacher effects

- ❖ Can we be confident that we've **fully accounted** for *other explanations* for test score gains, so that attribution to a specific teacher is appropriate?
- ❖ Given one year of test score gains, it is impossible to distinguish between teacher and *classroom* effects (more years helps)

(3) Isolating teacher effects

- ❖ Statistical models require **a lot of assumptions**, and results are often sensitive to these
 - Reasonable people can and do disagree about models
- ❖ Does attributing achievement gains to individual teachers even make sense?
 - Learning may not occur this way, especially in middle and high school settings

(3) Isolating teacher effects

- ❖ Suppose Mr. Arcilla teaches 7th grade ELA
- ❖ Mr. Arcilla's new history colleague, Mr. Zimmerman, is known for his ability to expand his students' vocabulary and comprehension of text
- ❖ If Mr. Arcilla's class has larger than average ELA score gains, does he have high "value-added?"

(4) Who counts?

- ❖ Not all teachers teach tested grades or subjects
- ❖ Not all students are tested
- ❖ Not all students contribute to value-added estimates
- ❖ Need to know students' *prior year* achievement
 - Thus only 4th – 8th grade possible
 - Missing for a large fraction of students (often 25%+)
- ❖ To whom do we attribute "mobile" students?
 - Mobility is extremely high in certain schools

(5) Precision

- ❖ All value-added measures are *estimates* based on a statistical model, and are subject to uncertainty: a **"margin of error"**
- ❖ Are teacher effects **precise enough to be useful?**
 - *Ranking* teachers requires pretty precise estimates
 - Easier to identify the "very best" or "very worst"
 - More years helps, but this may be "too late"

(5) Precision

2 My Results, Compared to Peer Teachers:

How do my results compare to other teachers in my grade and subject area throughout NYC, whose classrooms have similar predicted gains, adjusted for teacher experience levels?

- ↳ 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

	Number of Students	Average				Percentile (0-100%)	My Percentile				
		Prior Proficiency Rating	Actual Gain	Predicted Gain	Value-Added		0%	25%	50%	75%	100%
This year: 2007-08 Range	24	2.1	0.19	0.08	0.11	65% 46-84%	▼				
2006-07 Range	24	2.4	0.08	0.08	0.00	50% 30-70%	▼				
2005-06 Range	25	2.5	0.03	0.06	-0.03	43% 22-64%	▼				
Last 3 years average Range	73	2.4	0.10	0.07	0.03	53% 40-66%	▼				

(5) Precision

How to Interpret Teacher Data

The teacher shown in the chart below had 24 students in her class contribute to her Value-Added score in 2007-08. The average prior proficiency rating of these students was a 2.1. On average, these students gained 0.19 of a proficiency rating. On average, these students were predicted to gain 0.08 of a proficiency level by the Value-Added model. Thus, these students gained more than predicted. This teacher's "Value-Added score" is the difference between the actual gain and the predicted gain—in this case 0.11 ($0.19 - 0.08 = 0.11$). A Value-Added score of 0.11 puts this teacher in the 65th percentile, which means her Value-Added score is higher than 65% of the teacher's in the comparison group. While this teacher is most likely to be in the 65th percentile, we provide a range because there is some measurement uncertainty for all statistical calculations like this. For this teacher, we are 95% certain that she is between the 46th-84th percentile.

(6) Stability

- ❖ Because error is large in any one year, the **stability of teacher effect estimates from year to year** is quite low
- ❖ E.g. one study of Florida districts found:
 - Of teachers identified to be in the *bottom 20%* one year, **11-18% were in the top 20%** in the next
 - Of teachers identified to be in the *top 20%* one year, **10-14% were in the bottom 20%** in the next

(6) Stability

- ❖ Particularly a problem for estimates based on one year
- ❖ Estimates "**smooth out**" over time, but arguably this defeats the purpose of professional development needed now

Closing thoughts and questions

- ❖ Teacher quality is critical, and a better means for assessing teacher effectiveness is needed
- ❖ Existing research on value-added measurement suggests we should be **concerned about using** value-added to evaluate individual teachers
- ❖ We still have much to learn. For example...

Closing thoughts and questions

- ❖ How will teachers and school leaders use Teacher Data Reports in practice?
 - Research tells us virtually nothing about how value-added measures can be used to **improve instruction**
 - Value added measures are fundamentally only (estimated) **rankings**
 - NYC TDI will contribute to our understanding of how teachers and schools can and do use data
 - “Focus on student achievement” may be an end in itself

Closing thoughts and questions

- ❖ Do teacher effects on standardized tests have *long-run* consequences?
 - What we do know about value added measurement is based entirely on **short-run outcomes on high-stakes tests**—little validation against anything else
- ❖ Is value-added “better than the status quo?”
 - Some argue for “low-stakes” use of value-added
 - True? Possible? What are the alternatives?

References

- ❖ Some excellent (mostly non-technical) resources:
 - Koretz (2008) in *American Educator*
 - Braun (2005) primer for ETS
 - “Merit Pay for Florida Teachers: Design and Implementation Issues” (RAND 2007)
 - Rivkin (2007) CALDER policy brief