High-Stakes Teacher Evaluation: High Cost — Big Losses

As schools face public opinion plus state and federal policy changes, principals and teachers are evaluated partly on student performance. The advantages have been widely touted without serious attention to the policy consequences. One criticism: This may push schools to fire teachers based on poor student evaluations. The High-Stakes Teacher Evaluation Conference will present a thoughtful analysis with an eye toward policy issues and possible negative consequences.
Value-Added Models (VAMs) in the Context of Teacher Evaluation

Audrey Amrein-Beardsley
Opening Session
High-Stakes Teacher Evaluation: High Cost — Big Losses
University of Arizona
October 11, 2012
Introduction

The Socially Constructed Theory of Change:

• By holding districts, schools, teachers, and students accountable for meeting higher standards on high-stakes tests, administrators will supervise schools better, teachers will teach better, and as a result students will learn more and students’ test scores will increase.

• The theory of change is based on a paucity of empirical research evidence, however.

• Regardless, for the past 30 years (since A Nation at Risk) educational policymakers and politicians have continuously “reinvented” educational reform initiatives using the same theory of change.
Introduction to VAMs

• Accordingly, policy attention has now turned to new and improved measurement systems that “will” improve upon the accuracy with which we measure educator effectiveness.

• Value-added models (VAMs) are meant to isolate and measure teachers’ (or schools’/districts’) contributions to student achievement on large-scaled tests as groups of students move through school.

• VAM statisticians measure value-added by mathematically calculating the “value” a teacher (or school/district) “adds” to (or detracts from) student achievement scores from the point at which students enter classrooms to the points they leave.

• Again, however, limited empirical research evidence indicates that VAMs work in the ways theorized.
VAMBoozled

• We now have additional evidence indicating that this same theory of change, even with its advanced VAM metrics, is still flawed and misguided.

• The new and improved VAM metric systems are bamboozling, or in this case VAMBoozling many who are keenly interested in reforming America’s public school system.

• Those marketing and selling their VAMs, often for profit, are raking in billions in federal, state, and district monies.

• Not only are we being VAMBoozled, we are paying to be VAMBoozled as our political leaders use taxpayer revenues to further advance a false and highly deceptive, socially engineered theory of change.
Policy Context

• Rhetoric advancing the same theory of change is evident throughout No Child Left Behind (NCLB), the revisions of the Elementary and Secondary Education Act (NCLB “revised”), and, the federal Race to the Top initiative.

• What has changed?
  • Accountability shift from student-level focus to teacher-level focus
  • Waivers excusing states from not meeting NCLB’s “100% proficiency by 2014 goal,” if states agree to adopt even higher high-stakes for teachers
  • The ever-expanding role of the federal government and its control over state and local education policies and initiatives
  • Econometricians and value-added statisticians entering into the educational evaluation arena (for profit)
Policy History

- VAMs in econometrics (used to gauge marketplace inputs/outputs)
- VAMs introduced into education in late 1970s
- Post NCLB – Adequate Yearly Progress (AYP) issues
- Growth/VAM pilots in 2005 with NCLB revisions
- 2005-2010, 15 states received pilot funding from USDOE
- USDOE released report that growth pilots showed minimal impact on educational accountability
  - Different models cause dramatic differences
  - Varied levels of difficulty on state tests
  - Varied interpretations of results
  - Lack of stability
    - (i.e., one in three teachers (35%) would be misclassified with one year of data, one in four teachers (26%) would be misclassified with three years of data; and one in ten would be misclassified with 10 years of data).
To provide a current overview capturing the national landscape (i.e., post USDOE pilots and RttT)

- Growth and value-added systems include:
  - “Off the shelf” extant models, the most popular of which include the SAS Education Value-Added Assessment System (EVAAS) and the Value-Added Research Center (VARC) model.
  - The Student Growth Percentiles (SGP) Model (a.k.a., the Colorado Growth Model (CGM));
  - Homegrown models (e.g., value-table models; see, for example, Delaware);
  - Hybrid versions with systems incorporating components from various models.

- Existing reports (e.g., CCSSO, Education Commission of the States (ECS), etc.) are limited by data aggregation, regional reporting, or limited areas of focus (e.g., only legislation backing model implementation)

Citation:
Research Methods

- Phone interviews, email exchanges, state department of education websites
- Data collected June – December, 2011
- State representatives (e.g., Title IIA, Accountability & Assessment, Leadership & Evaluation, and Teacher Quality or Leadership Departments and Divisions)
- 50 states and D.C.
Questionnaire

1. State legislation requiring that the evaluation of teacher effectiveness integrate student growth/VAM data
2. Growth or value-added model used
3. Test types and grade levels included
4. Whether models account or control for student demographics
5. Statewide data systems in place to link students with teachers
6. Teacher consequences attached to growth/VAM output
7. Percent of student growth/VA required for teacher evaluation
8. Teachers using data in formative ways
9. Strengths and weaknesses/limitations with models being implemented/adopted
Growth v. Value-Added Models

- 20 states + D.C. (40%) have already implemented growth/VAMs

- Growth models (12 states) 24% (i.e., SGP)
  - Estimate school or teacher impact on student achievement
  - Typically used for more descriptive purposes
  - Measure students’ growth in relation to similar students

- Value-Added Models (8 states + D.C.) 18% (e.g., EVAAS)
  - Estimate school or teacher impact on student achievement
  - More likely to be used for consequential purposes
  - Blocks built into models to control for the effects of external factors
Growth v. Value-Added Use

- SGP (12) 24%
- VAM (8 + D.C.) 17%
- Pilot/Dev (18) 35%
- No plans (7) 14%
- Local control (3) 6%
- Value Table (1) 2%
- Hybrid (1) 2%

*AK not shown, no plans
State Legislation

Legislation Requiring Teacher Evaluations using Student Growth/VAM Output

30 states + D.C. (61%)
Descriptive Findings

• 34 states and D.C. (67%) have data systems in place to link individual students with teachers.
  • 12 states (24%) are piloting/still developing data systems.
  • 4 states (8%) do not have such data systems or plans to develop them.

• 100% of the states currently using growth or VA models use their state large-scale standardized tests in mathematics and English/language arts to generate accountability data.

• Nine states (18%) are evaluating (or are planning to evaluate) high school teacher effectiveness using growth/VAMs (e.g., using end-of-course exams)

• Two states (4%) are evaluating (or are planning to evaluate) primary teacher effectiveness (e.g., using the Northwest Education Association Measures of Academic Progress (NWEA MAP))
Consequential Findings

At the state level...

- 15 states (29%) are still determining what teacher-level consequences they will tie to growth/VAM data.
- 13 states and D.C. (27%) are using (plan to use) growth/VAM data to identify professional development needs.
- 10 states and D.C. (22%) are using (plan to use) growth/VAM data to make teacher tenure decisions.
- 9 states and D.C. (20%) are using (plan to use) growth/VAM data to make teacher termination decisions.
- 9 states and D.C. (20%) are using (plan to use) growth/VAM data to determine merit pay, bonus, or pay for performance decisions.
Notable Concerns

• Reliability
  • Fluctuations and year-to-year misclassifications

• Validity
  • Design/intent of state tests for such purposes
  • Data contradictions with other teacher evaluation tools (e.g., low correlations)

• Fairness
  • Inability to assess teachers of non-tested subject areas and teachers in ineligible grades

• Transparency
  • Difficult to explain and understand (e.g., EVAAS model)

• Formative Use
  • Not one state representative was able to articulate a statewide plan for formative data use
Houston – Research Project 2

• To understand how teachers in the Houston Independent School District (HISD) have been impacted by EVAAS implementation and use for the past five years

• BlueDream.tv Video: EVAAS, Value-Added and Teacher Branding - Click here

Citation:


See also:


Houston – Research Project 3

To further understand the intended and unintended consequences of EVAAS implementation and use, as experienced and explained by HISD teachers.

- Reliability
- Validity (content- and criterion-related)
- Fairness
- Transparency
- Formative use
- Intended consequences and claimed benefits of EVAAS
- Unintended consequences

Citation:

Research Methods

- Mixed-methods, survey research study
- Questions formulated using conceptual framework as aligned with educational measurement standards (AERA, APA & NCME, 1999)
  - Reliability
  - Validity (content- and criterion-related evidence)
  - Fairness
  - Transparency
  - Formative use
  - Intended consequences and claimed benefits of EVAAS
  - Unintended consequences
Response Rate

• Response rate (882/≈3,900) 23.0%

• Quantitative Data
  – Chi-square analyses for union vs. non-union teachers

• Qualitative Data
  – Open-ended, free response questions generated 4,594 unique responses
  – Teacher quotations to demonstrate lived experiences (Creswell, 2003)
  – Triangulation using qualitative and quantitative survey data
Findings: Reliability

- Consistent EVAAS scores (42%) and inconsistent EVAAS scores (46%)
  - Inconsistencies were an issue half of the time
    - “In three years, I was above average, below average and [then] average.”

- Similar inconsistencies were reported by teachers who taught across grade levels and subject areas

- Teachers attributed inconsistencies to the types of students they taught (i.e., bias)
  - ELL, transitional, special education, gifted students
  - Teachers also attributed such fluctuations to the nonrandom assignment of students into their classrooms
    - “[EVAAS] depends a lot on home support, background knowledge, current family situation, lack of sleep, whether parents are at home, in jail, etc. [There are too many outside factors – behavior issues, etc.”
Findings: Content-Related Validity

• Data imperfections (10%)
  – Fractional responsibilities and the allocation of instructional time
  – Team-teaching, departmentalized, lab teachers, multi-grade teachers

• Not teachers of record for students (18%)
  – Student mobility, late additions and early leavers, alternative school students or students in “pull out” programs

  • “I'm not sure how I get evaluated for a student who is only in my class for one month and then goes into CEP [community education partners for disciplinary alternative education]. I'm still considered the teacher of record even though he spent 5-6 months out of my classroom.”
Findings: Criterion-Related Validity

• Inconsistencies between EVAAS and principal evaluation scores (50%)
  – Also noted, principals adjusting scores to match EVAAS

• Inconsistencies given teacher awards and recognitions, and also teacher-leader positions (32%)
  – “I have always received positive - even glowing - observation and evaluation scores from my principal and evaluator. I have been asked to serve as a lead teacher on campus and I have mentored others - but my negative [EVAAS] growth score does not reflect that.”
Findings: Fairness

• Unfair teacher evaluation tool (19%)
  – Not all teachers were included in the sample, so likely an underestimate

• Particularly unfair if used for consequential decision-making
  – Rewards teachers who use questionable/gaming practices
  – Different students influence teacher EVAAS scores

• Other issues with fairness
  – Too many factors bias results: external factors, curricula and test (mis)alignment, certain student populations, etc.
Findings: Transparency

Interpreting the Teacher Value-Added Report

Use this report to evaluate how well a teacher facilitates student progress. The district gain is an estimate of the district’s influence on student progress. The Teacher Value-Added Report compares each teacher’s gain to the district gain. This comparison indicates how a teacher influences student progress in the given subject.

**Teacher NCE Gain, Standard Error**
The Teacher NCE Gain is a conservative estimate of a teacher’s influence on students’ academic progress estimated by using all students linked to a teacher who were tested on TAKS, TAKS Accommodated, or Stanford/Other in non-TAKS grades and subjects. It is expressed in state NCEs using 2006-2008 as the base year. The Tch Std Error provides the basis for establishing a confidence band around the Teacher NCE Gain value. One Standard Error is used in the statistical test reported under Teacher Comparison to the HISD Ref Gain. Note that this year’s estimate of previous year’s gains may have changed as a result of incorporating the most recent student data.

**HISD Reference Gain**
The HISD Reference Gain is the gain made by HISD in this subject and grade. This gain is expressed as an NCE, based on the state population in 2006. A positive gain indicates HISD made more progress than the state average and a negative gain indicates HISD made less progress than the state average.

**Teacher Comparison**
The Teacher Comparison to HISD Ref Gain column shows whether there is a difference in the progress rate for the teacher compared to the HISD Reference Gain. Comparisons are based on one standard error.

- Above means that students taught by this teacher made more progress than the reference gain. A teacher classified as "above" will have a Gain Index greater than 1.
- Below means that students taught by this teacher made less progress than the reference gain. A teacher classified as "below" will have a Gain Index less than 1.
- NDD means that the progress made by the teacher’s students was not detectably different from the reference gain. A teacher classified as "NDD" will have a Gain Index between -1 and 1.

**Teacher Gain Index**
To calculate the Teacher Gain Index, first subtract the HISD Reference Gain from the Teacher NCE Gain. Then this difference is divided by the Tch Std Error. The 2010 Teacher Gain Index is used by HISD for determining ASPIRE Awards.

---

Interpreting the Report for Teacher Reflection

Use this report to identify patterns or trends of progress among groups of students at different achievement levels. This report is intended for diagnostic purposes only, since students with missing scores in the current or previous year are excluded. Therefore, you will not be able to calculate Teacher NCE Gain from the table above.

**The Chart**
The chart offers a visual representation of average student progress for students at different entering achievement levels. The green line is the district reference gain line, representing the average amount of progress made by students in HISD. Blue bars show the progress of students in the current school year. Bars below the green line indicate that students in the subgroup made less progress than the district average. A confidence interval of one standard error is indicated by the red bar. No bar is presented for subgroups with fewer than five students.

**Student Assignment**
Prior-achievement subgroups are determined by averaging a student’s NCE score in the current year with his or her NCE score in the previous year. Students are assigned to subgroups based upon where they fall in the distribution of all students in the district who took the test in the years included in the analysis. As a result, some teachers may find that certain subgroups are more heavily populated than others.

**The Table**
The table immediately below the chart allows for a comparison of the subgroup NCE gains to the HISD reference gain. The standard error is reported immediately below the Average NCE Gain for each subgroup. The row labeled Nr of Students shows the number of students in the subgroup. When there are fewer than five students in a subgroup, the number of students in the group is reported, but the gain is not displayed.
Findings: Transparency

- 40% of teachers reported using or attempting to use EVAAS but called the reports “vague” and “unclear.”

- Teachers were “not quite sure how to interpret” or use EVAAS data to inform or improve upon their classroom instruction.
  - “EVAAS is most helpful for me when looking at subgroups and their growth. For example, you can look at the growth of just boys, or girls, or gifted or non-gifted students in your class...The only issue is that you're not 100% sure how this score is calculated, so it's not clear what part of your practice you should go back and change.”

- Teachers reported not understanding EVAAS output, and related, principals not discussing their EVAAS data with them.
Findings: Formative Use

- 60% of teachers reported not using EVAAS data whatsoever.
- Teachers reported district-wide discrepancies given the release of EVAAS reports.
- HISD lacks a cohesive district-wide plan for the distribution of EVAAS data, as well as a plan for formative use.
  - Again, 40% of teachers reported using or attempting to use EVAAS but called the reports “vague” and “unclear.”
  - Teachers questioned whether formative use was even possible.
Findings: Intended Consequences

• “EVAAS reports are simple to use.”
  – 67% disagreed.

• “EVAAS helps you become a more effective teacher.”
  – 65% disagreed.

• “EVAAS helps increase student learning.”
  – 63% disagreed.

• “EVAAS helps improve instruction.”
  – 59% disagreed.

• The majority of respondents reported not believing that the EVAAS works in the ways intended and as advertised.
Findings: Unintended Consequences

- Switching grade levels or subject areas biases EVAAS scores
- Teaching students back-to-back (i.e., looping) makes it difficult to post high value-added scores two years in a row
- Lowered morale and heightened competition
- Teachers learning how to game the system
  - Teaching to the test
  - Teachers identifying students who they would both select and avoid in order to achieve the highest possible EVAAS scores
- High-stakes EVAAS use is exacerbating unintended effects.
  - “Teachers have become even more distrustful of each other because they are afraid that someone might steal a good teaching method or materials from them and in turn earn more bonus money...it actually is detrimental to students because teachers are not willing to share ideas or materials that might help increase student learning and achievement.”
Overall Conclusions

- Unintended consequences will accompany the intended consequences of implementing any value-added model, especially when high-stakes decisions are to be made using growth/VAM output.
  - Campbell’s Law – overreliance on growth or value-added data will inevitably corrupt the educational process.

- Growth/Value-added data should not [yet, if ever] be used for high-stakes decision-making, and only perhaps for professional development, formative, and large-scale evaluative purposes [if that].
  - Whether any of these activities are possible using growth or VAM output is still open for debate. More research is certainly needed here.
Overall Conclusions Cont.

• Teachers’ behaviors, perspectives, and opinions need to be continuously examined as they are on the ground levels, experiencing what growth/VAMs do, when theory is translated into practice.

• We must better understand whether “sophisticated” statistical controls and blocks can mediate the effects of the nonrandom assignment of students into classrooms.

• We must better understand whether using multiple measures along with growth or VAM data, in line with our professional measurement standards, will [ever] work.

• We MUST do a better job translating research to practice.
  – E.g., user-friendly videos, blogs, policy briefs, other “high-impact” and highly accessible scholarship
High-Stakes Teacher Evaluation: High Cost — Big Losses

University of Arizona
October 12, 2012
Why do policy makers believe that teacher effects should be stable?

Thomas Good will discuss classroom complexities and whether stability can be expected of teacher effects over time. Given that some teachers are consistent in their effects on student achievement over time, he will pursue information on how "more effective" teachers teach. Tom will then discuss data limitations on explaining how teachers influence student achievement. Tom will conclude with a few thoughts on why our knowledge base is so limited in this area.
What are teacher effects anyways, and how can research in this area contribute to our thinking about whether causal statements can be made about them?

Spyros Konstantopoulos will provide a literature review about the magnitude of teacher effects (including some research on VAMs) and whether the effects vary by different levels of education or by different subject matter. Spyros will also discuss what is known about the stability of teacher effects over time.
How might supervisor observational scores help us to understand and measure teacher effectiveness, particularly within such accountability systems, and as aligned with our educational measurement standards (AERA, APA & NCME, 1999)?

Heather Hill and Corinne Herlihy will examine how state and local entities intend to ensure that teacher observation systems provide valid and reliable scores. Through a document analysis and interviews with state and local education officials, the authors explore several issues with observational systems, including the overall generalizability of teacher scores; the training, certification, and reliability of observers; and specifications regarding the sampling and number of lessons observed per teacher.
Teacher Education Accountability – Rick Ginsberg and Neal Kingston, University of Kansas

What are the implications of such teacher evaluation systems for higher education, namely the teacher education programs that are also coming under increased scrutiny to also demonstrate their “value-added?”

Rick Ginsberg and Neal Kingston will explore accreditation in teacher education, the situation facing teacher education, and, related, the use of outcome measures in a variety of other professional fields. As well, they will assess the strengths and weaknesses of the potential measurement tools being considered, the iatrogenic problems that invariably result from the forced implementation of inappropriate accountability treatments, and some possible avenues for the field to consider in dealing with the reality of today's accountability-driven atmosphere.
Tomorrow’s “Value-Added”

Over-Simplifying Complex Systems – David C. Berliner, Arizona State University

Are the intended consequences of these systems attainable, given the myriad of methodological issues such sophisticated systems cannot [and may not ever be able to] handle?

David Berliner will discuss how existing analytic models of teacher effectiveness can examine only a few variables simultaneously, leaving a myriad of unobserved variables interacting with each other and making every school year and every class taught different from another. Unobserved variables are countless and thus have something in common with chaos theory, namely, that the conditions that are unobserved, even if quite subtle and even quite far from the classroom, can still determine achievement outcomes of a classroom.
Assumptions and Unintended Effects – Alyson L. Lavigne, Roosevelt University

Will firing ineffective teachers improve student achievement? What are some other unanswered questions, unchecked assumptions, and unintended effects?

Alyson Lavigne will explore the logic behind growth and value-added models, as situated within high-stakes teacher evaluation systems. Alyson will also put into check the assumptions and rationales on which such systems are based. For example, will firing ineffective (as primarily measured by observation and value-added scores) teachers will improve student achievement? Alyson will also examine how high-stakes teacher evaluation decisions are changing and could change the landscape of American education.
Tomorrow’s “Value-Added”

Closing Panel, featuring:

• Ronald W. Marx, University of Arizona*
• Deb Duvall, Arizona School Administrators
• Andrew Morrill, Arizona Education Association
• Jane West, American Association of Colleges for Teacher Education

Panelists will speak about the conference sessions, representing the various constituencies they represent (i.e., academics, teachers, superintendents, principles, and state level policy actors). Panelists will also discuss overall concerns, issues, and resolutions.
Enjoy!

&

Thank You!

Audrey Amrein-Beardsley
audrey.beardsley@asu.edu
As schools face public opinion plus state and federal policy changes, principals and teachers are evaluated partly on student performance. The advantages have been widely touted without serious attention to the policy consequences. One criticism: This may push schools to fire teachers based on poor student evaluations. The High-Stakes Teacher Evaluation Conference will present a thoughtful analysis with an eye toward policy issues and possible negative consequences.
Americans are not and never have been satisfied with teachers, schools and student performance.

Why is that?
How to Identify Effective Teachers?
30 states are now evaluating teachers on the basis of student achievement

Why - because teachers impact student achievement

- It is widely believed that student achievement is too low;
- So when a teacher's score is low what do you do?
- Fire or help?
- Firing is a risky strategy because it assumes better replacements are readily available;
- Improvement assumes a knowledge base;
- What do teachers do to increase achievement?
# The Core Problem in Five Major Reforms

<table>
<thead>
<tr>
<th>Study</th>
<th>Category</th>
<th>Suggested Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sputnik Launch/ New Math 1957</td>
<td>CURRICULUM</td>
<td>Math/science needed major reform.</td>
</tr>
<tr>
<td>A Nation at Risk 1983</td>
<td>STUDENTS</td>
<td>Students need to work harder.</td>
</tr>
<tr>
<td>Prisoners of Time 1994</td>
<td>CORE SUBJECTS NEGLECTED</td>
<td>Allocate more time to core subjects and use more ambitious strategies.</td>
</tr>
<tr>
<td>No Child Left Behind 2001</td>
<td>SCHOOLS</td>
<td>Identify failing schools – correct or close.</td>
</tr>
<tr>
<td>Race to the Top 2011</td>
<td>STATES and TEACHERS</td>
<td>Provide states with incentives to encourage high stakes evaluation.</td>
</tr>
</tbody>
</table>

The effects of teaching on students’ achievement: What we know and why we know so little

Four questions:

• Why are Americans dissatisfied with teaching and teachers?

• Why is teaching seen as easy?

• How do teachers influence student achievement?

• Why is our knowledge base so limited?
Is There Anything Teachers Are Not Responsible For?
What Should Teachers Know?

- *Teachers should*
  - Know the subject well.
  - Convey subject matter in accessible and accurate ways.
  - Assess degree of student subject matter learning.
  - Help students to apply content knowledge to problems.
  - Help students to use content knowledge to find new problems.
  - Help students to value subject matter.
They Should Also...

- Respect student diversity and teach students to value diversity.
- Promote citizens for tomorrow who value democracy.
- Promote self-expression.
- Teach students to be patient and take turns.
- Teach students to initiate and to lead.
And...

• Create safe environments and handle disciplining.
• Create pleasant environments for students.
• Work in teams to improve curriculum.
• Work in teams to improve school governance and functioning.
• Share ideas and mentor other teachers.
Most citizens believe that teaching is easy

- Innovations
- Goals 2000
- DC schools’ plans for radically improving teachers
Enthusiastically Recommend Innovations: Some Examples from the Past
Consider these:

- Tape Recorder
- TV
- Internet
- Chalk
- Blackboard
- Team Teaching
- Expanded Class Period
- Thematic Teaching
- Teaching Combinations – Teach Math/Science Together
- Principal as Leader
Consider these:

- School Wide Reform
- Looping
- School Leadership Teams
- Merit Pay
- Career Ladder
- Summer School
- Pencil Sharpener
- Peer Tutoring
- Project Based Learning
- Mastery Learning
- Individualized Learning
- Phonetics
Consider these:

• Whole Language
• Grade Retention
• Grade Promotion
• Magnetic Schools/Charter Schools
• Gifted Instruction
• Cross-grade Tutoring
• Small Group Learning
• Pair-share
• Competition
• Cooperation
• Charter Schools
• Voucher Plans
Consider these:

- Choral Responding
- Seatwork
- Immersion
- Distributed practice
- Ability Grouping
- Heterogeneous Grouping Objectives
- Behavioral
- Advanced Organizers
- Overhead projector
- Mimeograph machine
- Xerox copier
- Schools within schools
- Magnet schools
National Education Goals

By the year 2000:

1. School Readiness: All children in America will start school ready to learn.
2. School Completion: The high school graduation rate will increase to at least 90%.
3. Mathematics and Science: U.S. students will be first in the world in science and mathematics achievement.

DPCS Wins Over $62 Million in Federal Grant to Reward and Retain Great Teachers

- What did District of Columbia Public Schools promise to get this grant?
- At least 90 percent of DPCS teachers and principals will be Highly Effective or Effective, as determined through the IMPACT evaluation system.

2012
Were Americans Concerned About Education in 1900?
What Grade Would You Give Public Schools Nationally?*

<table>
<thead>
<tr>
<th></th>
<th>2009 (%)</th>
<th>2008 (%)</th>
<th>2007 (%)</th>
<th>2006 (%)</th>
<th>2005 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A &amp; B</td>
<td>19</td>
<td>22</td>
<td>16</td>
<td>21</td>
<td>24</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>18</td>
<td>19</td>
<td>14</td>
<td>19</td>
<td>22</td>
</tr>
<tr>
<td>C</td>
<td>55</td>
<td>44</td>
<td>57</td>
<td>51</td>
<td>46</td>
</tr>
<tr>
<td>D</td>
<td>19</td>
<td>13</td>
<td>18</td>
<td>14</td>
<td>13</td>
</tr>
<tr>
<td>Fail</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Don’t Know</td>
<td>1</td>
<td>16</td>
<td>4</td>
<td>11</td>
<td>13</td>
</tr>
</tbody>
</table>

*Question reworded here for simplicity.

*(Bushaw & McNee, September 2009)
What Grade Would You Give the School Your Oldest Child Attends?*

*Question reworded here for simplicity.*

*(Bushaw & McNee, September 2009)*

<table>
<thead>
<tr>
<th></th>
<th>2009 (%)</th>
<th>2008 (%)</th>
<th>2007 (%)</th>
<th>2006 (%)</th>
<th>2005 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A &amp; B</td>
<td>74</td>
<td>72</td>
<td>67</td>
<td>64</td>
<td>69</td>
</tr>
<tr>
<td>A</td>
<td>31</td>
<td>30</td>
<td>19</td>
<td>26</td>
<td>31</td>
</tr>
<tr>
<td>B</td>
<td>43</td>
<td>42</td>
<td>48</td>
<td>38</td>
<td>38</td>
</tr>
<tr>
<td>C</td>
<td>17</td>
<td>14</td>
<td>24</td>
<td>24</td>
<td>21</td>
</tr>
<tr>
<td>D</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Fail</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Don’t Know</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>
Despite favorable parents’ attitude toward teachers, teachers are under constant criticism

“Teachers ignore, resist, subvert, misinterpret, selectively adopt, or otherwise distort reformers’ intentions. Changes tend to be superficial, seldom penetrating the core of instructional practice”

Lefstein (2008)
Teachers’ effects on students achievement from year to year

How does their stability compare to
• professional bowlers?
• professional golfers?
• college basketball free-throw shooting?
Bowling anyone?
The conditions for bowling are clear as well as the rules for scoring

- The playing surface is a lane 60 feet long, 42 inches wide;
- The pin is made of Hard Maple, it has a diameter of 2.25 at the base and a circumference of 15 inches at its widest point; the pin must weigh between 3 pounds 6 ounces and 3 pounds 10 ounces;
- The ball has a circumference of no more than 27 inches, has to weigh from 10 to 16 pounds and may have 2 or 3 finger-holes.

Bowler’s performance is scored with 100% reliability
# Bowling Ranking

<table>
<thead>
<tr>
<th></th>
<th>Name</th>
<th>2008-2009</th>
<th>2009-2010</th>
<th>2010-2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Norm Duke</td>
<td>1</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>Rhino Page</td>
<td>2</td>
<td>9</td>
<td>NR</td>
</tr>
<tr>
<td>3</td>
<td>Mike Scroggins</td>
<td>3</td>
<td>3</td>
<td>NR</td>
</tr>
<tr>
<td>4</td>
<td>Patrick Allen</td>
<td>4</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>5</td>
<td>Wes Malott</td>
<td>5</td>
<td>7</td>
<td>NR</td>
</tr>
<tr>
<td>6</td>
<td>Chris Barnes</td>
<td>6</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>Walter Ray Williams Jr.</td>
<td>7</td>
<td>1</td>
<td>NR</td>
</tr>
<tr>
<td>8</td>
<td>Bill O’Neill</td>
<td>8</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>Brad Angelo</td>
<td>9</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>10</td>
<td>Mike Fagan</td>
<td>10</td>
<td>8</td>
<td>17</td>
</tr>
</tbody>
</table>
# Bowling Ranking

<table>
<thead>
<tr>
<th></th>
<th>Name</th>
<th>2008-2009</th>
<th>2009-2010</th>
<th>2010-2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>Pete Weber</td>
<td>11</td>
<td>10</td>
<td>NR</td>
</tr>
<tr>
<td>12</td>
<td>Parker Bohn, III</td>
<td>12</td>
<td>NR</td>
<td>18</td>
</tr>
<tr>
<td>13</td>
<td>John Nolen</td>
<td>13</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>14</td>
<td>Tommy Jones</td>
<td>14</td>
<td>5</td>
<td>16</td>
</tr>
<tr>
<td>15</td>
<td>Mika Koivuniemi</td>
<td>15</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>Danny Wiseman</td>
<td>16</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>17</td>
<td>Steve Harman</td>
<td>17</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>18</td>
<td>Sean Rash</td>
<td>18</td>
<td>18</td>
<td>7</td>
</tr>
<tr>
<td>19</td>
<td>Jeff Carter</td>
<td>19</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>20</td>
<td>Ryan Shafer</td>
<td>20</td>
<td>NR</td>
<td>NR</td>
</tr>
</tbody>
</table>
# Best Free-throw College Basketball Teams

<table>
<thead>
<tr>
<th>2009-2010</th>
<th>2010-2011</th>
<th>2011-2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Brigham Young</td>
<td>15</td>
<td>NR</td>
</tr>
<tr>
<td>2. Indiana State</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>3. U Colorado</td>
<td>5</td>
<td>NR</td>
</tr>
<tr>
<td>3. Harvard</td>
<td>2</td>
<td>NR</td>
</tr>
<tr>
<td>5. Lafayette</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>5. St. Mary’s</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>7. Utah State</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>8. Texas – San Antonio</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>9. Duke</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>10. Drake</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>10. California – Berkley</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>12. Northern Iowa</td>
<td>18</td>
<td>NR</td>
</tr>
<tr>
<td>13. Villanova</td>
<td>17</td>
<td>20</td>
</tr>
<tr>
<td>14. Pennsylvania State</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>14. Cleveland State</td>
<td>NR</td>
<td>NR</td>
</tr>
</tbody>
</table>
Sean Miller 2012 University of Arizona head basketball coach won the “Shots from the Heart” free-throw charity event.

How did the University of Arizona basketball team rank nationally?

<table>
<thead>
<tr>
<th>Year Range</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009-2010</td>
<td>30</td>
</tr>
<tr>
<td>2010-2011</td>
<td>32</td>
</tr>
<tr>
<td>2011-2012</td>
<td>110</td>
</tr>
</tbody>
</table>
Golf anyone?
## World Golf Ranking

<table>
<thead>
<tr>
<th>Year</th>
<th>2007</th>
<th>2008</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>3&lt;sup&gt;rd&lt;/sup&gt;</td>
</tr>
<tr>
<td>1&lt;sup&gt;st&lt;/sup&gt; 50</td>
<td>37</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>2&lt;sup&gt;nd&lt;/sup&gt; 50</td>
<td>8</td>
<td>17</td>
<td>9</td>
</tr>
<tr>
<td>3&lt;sup&gt;rd&lt;/sup&gt; 50</td>
<td>3</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>4&lt;sup&gt;th&lt;/sup&gt; 50</td>
<td>1</td>
<td>7</td>
<td>7</td>
</tr>
</tbody>
</table>
World Golf Ranking

2007 - 2008

Out of best 200 golfers 48 were not on the top list next year.

2007 - 2010

Out of best 200 golfers 88 were not on the top list in three years.
Teaching vs. Golf Performance

• In golf performance varies due to weather, health, player’s mood, etc., just as does teacher’s performance;
• YET, conditions of golfing are much more stable than in case of teaching:
  - same tee to green distance;
  - standard equipment;
  - personal choice of a caddy.
• YET, conditions of teaching are notably unstable.
  - Uneven resources;
  - Students’ turnover;
  - Student absences and extended illnesses.
Do we have a knowledge base linking teachers instructional behavior to students’ gains on standardized tests?
Back to the Future
In 1970 teachers were not seen as critical influences on student achievement

Student achievement was believed to be predicted primarily by family circumstances and by heredity. This belief suggested a presumed ineffectiveness of teachers.

These beliefs were presumably empirically supported by the common report.

Good, Biddle and Brophy refuted these beliefs by logical argument and empirical data.
“Just as data suggesting that teachers do not make a difference frequently have been overgeneralized and accepted uncritically, there is a danger that these same mistakes can be made with the kinds of data cited in this chapter. While these data do establish that certain teachers consistently outperform others ... the present state of research in the area does not allow for the use of process-product findings for accountability purposes”.

(Good, Biddle and Brophy, Teachers Make a Difference, 1975)
Missouri Mathematics Naturalistic Research

• Identified teachers who were consistently high or low in their impact on student’s mathematics achievement on the Iowa Test of Basic Skills;
• Used existing research to design a coding system for describing teachers’ instruction during mathematics;
• Made classroom observations to see how high and low achieving teachers taught;
• Aggregated these discrete behaviors into an instructional approach;
• Designed manuals and a workshop for conveying this instructional approach.
Missouri Mathematics Experimental Study

- Obtained a school district that was willing to implement the study;
- Teachers were given information about the project and why we thought it might work;
- Teachers were assigned to treatment or delayed information group;
- All teachers were observed;
- Teachers generally implemented the program;
- Experimental teachers outperformed delayed information teachers;
- Although there was a main effect, it was also the case that some combinations of teachers and students showed more or less achievement gain.
These results were known in 1975

When Gerald Ford was the President
Who Cooks the Meal Is the Key.
Good Cooking and Good Teaching Have Similarities.

“As we know from our own experience, a chicken dinner with salad, wine, and apple pie can be a completely different experience as we move from restaurant to restaurant or eat at different homes. Please understand that I would like to add to this repertoire of choices: I say yes to better wine, and I support good appetizers and even organic vegetables. However, understand that the literature on effective teaching is not established on evidence showing that effective teachers bring in new components or better ingredients. Rather, the literature indicates that some teachers work with basic ingredients better than do other teachers.”

—Good (2009), p. 48
State Assessment Circa 1975

In 1975 Good, Biddle and Brophy noted that Beers and Campbell (1973) aptly summarized our feelings about the new movement toward state assessment. “It is probably safe to say that statewide assessment will not produce any startling revelations about what can be done by teachers with pupils to help children learn more effectively. This conclusion is not meant to be as much an indictment of statewide assessment as it is a statement of its limitations.”

Oddly this literature base linking classroom instruction and student achievement largely disappeared from the literature only to be rediscovered

- Strong (2011), Highly Qualified Teachers
- Ripley (2010), What Makes A Great Teacher
- Farr (2011), Teaching As Leadership
- Hattie (2009), Visible Learning
Strong, 2011

- Lesson objective is clearly expressed to students;
- Teacher understands student background and prior knowledge;
- Teacher and/or students demonstrate consideration of the topic in more than one way;
- Teacher provides/elicits from students multiple examples;
- Teacher uses appropriate non-examples;
- The pace of the lesson is not too slow or too fast;
- Feedback to students is about the process – not just correctness;
- Teacher uses guided practice – students practice and immediate feedback is provided;
- Teacher explains the Math concepts with good clarity;
- Teachers are actively engaged – most time spent on task.
Good Teaching:
Based on Amanda Ripley’s 2010 analysis of Teach for America Data

- Set big goals for their students;
- Constantly reevaluated what they are doing;
- Used ways to involve kids in the lesson (e.g. Mental Math);
- Demonstrated the learning that students were to engage in;
- Teachers spent time explaining the information;
- Implemented an “I do, we do, you do”; 
- Built up well established routines.
Teach for America:
Farr 2010

- Communicating key ideas;
- Coordinating student practice;
- Checking for student understanding;
- Tracking progress;
- Using organization and routine to maximize efficiency;
- Asserting authority by consistently following through on high expectations.
Visible Learning
Hattie, 2009

- Teachers powerfully impact student learning;
- Teachers need to be directive and actively engaged in teaching and learning;
- Teachers need to know how students are constructing meaning;
- Teachers need knowledge of the content they teach in order to provide meaningful feedback to students;
- Teachers need to be clear about intended meaning of lessons and making these success criteria explicit;
- Teachers must recognize that the learner’s construction of this knowledge that is critical;
- Teachers need to recognize and welcome student error and creating environments where students feel safe to learn and re-learn

Based on material presented by Hattie in Visible Learning, pp.238-239
Good and Grouws

1975

• Provides a review of material previously presented, collect homework, and ask students to engage in mental computation work;
• Teacher conducts development lesson where the teacher focusses on meaning and promoting student understanding through discussion and demonstration;
• Development also includes the assessment of student understanding the asking of both process and product questions and controlled practiced where teachers elaborate upon the meaning as necessary;
• Conduct homework using appropriate pace and alerting and accountability;
Good and Grouws

1975

• Students are given 15-20 minutes of uninterrupted time for successful practice;
• Seat work involves momentum – keeping the ball rolling – get everyone involved;
• Seat work also involves appropriate alerting (e.g. their work will be checked & accountability, the work is actually checked);
• Homework assignment if students have done seat work well;
• Homework should include review problems.
What have we learned about teaching that improves student achievement?
General Principles of Effective Teaching

1. Appropriate Teacher Expectations
2. Effective Use of Time
3. Proactive Classroom Management
4. Supportive and Caring Classrooms
5. Opportunity to Learn
6. Coherent Curriculum Content
7. Curriculum Alignment
8. Thoughtful Discourse
9. Scaffolding Students’ Ideas and Task Involvement
10. Practice/Application
11. Goal Oriented Assessments

3 examples of this knowledge base and the difficulty in using it

• Appropriate teacher expectations;
• Appropriate use of instructional time;
• Good classroom management.
Teachers’ expectations can influence student achievement.
What Do High-Achieving Students Receive?

• More opportunity to respond to teacher questions
• More opportunity for follow-up questions – tell me more
• More time to respond to teacher questions
• More *stay-with* behavior
• More teacher praise
• More opportunity for choice
What Do Low-Achieving Students Receive?

- Less opportunity to respond to teacher questions
- Less opportunity for follow-up questions
- Less time to respond to teacher questions
- More *give-up* behavior
- More teacher criticism
- Less opportunity for choice
• Expectations can be too low
• Expectations can be too high
• Inappropriate expectations can be expressed toward individual Students
• Inappropriate expectations can be expressed to the class as a whole
• Classroom Observation Systems DO NOT measure all of these dimensions
Teacher’s effective use of time can influence students’ achievement
• Teachers vary in their ability to manage time
• Information about variation in time management has been known for a long time
• Teachers often have difficulty in starting and ending lessons
• Teachers’ transitions from one part of a lesson to another are often time-consuming
Description of Fifth Grade Classroom Settings: Mean Number of Intervals in which Time-Sampled Codes Were Observed

<table>
<thead>
<tr>
<th>Category</th>
<th># of intervals</th>
<th>S.D.</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literacy/Language arts</td>
<td>21.97</td>
<td>8.64</td>
<td>0 – 48.00</td>
</tr>
<tr>
<td>Mathematics</td>
<td>14.64</td>
<td>6.45</td>
<td>0 – 42.75</td>
</tr>
<tr>
<td>Transition/Management + Managerial instructions</td>
<td>21.03</td>
<td>5.50</td>
<td>0 – 39.00</td>
</tr>
</tbody>
</table>
Prisoner of Time Report: Hours Spent on Basic Math, Science, and History

<table>
<thead>
<tr>
<th>Country</th>
<th>U.S.</th>
<th>Japan</th>
<th>France</th>
<th>Germany</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hours</td>
<td>1,460</td>
<td>3,170</td>
<td>3,280</td>
<td>3,528</td>
</tr>
</tbody>
</table>

Graph shows the number of hours spent on basic math, science, and history in the U.S., Japan, France, and Germany.
# Teachers Use of Time in Three Countries

<table>
<thead>
<tr>
<th></th>
<th>United States</th>
<th>Japan</th>
<th>Taiwan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher work with whole class</td>
<td>46%</td>
<td>86%</td>
<td>77%</td>
</tr>
<tr>
<td>Teacher provides math info</td>
<td>25%</td>
<td>33%</td>
<td>63%</td>
</tr>
<tr>
<td><strong>Student Engagement</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 1</td>
<td>69.8%</td>
<td>79.2%</td>
<td>85.1%</td>
</tr>
<tr>
<td>Grade 5</td>
<td>64.5%</td>
<td>87.4%</td>
<td>91.5%</td>
</tr>
</tbody>
</table>

Teacher’s appropriate management skills can increase students’ achievement
Effective management is related to students’ achievement

Some of the key-management aspects are:

• Withitness
• Overlapping
• Alerting
• Accountability
• Smoothness
Why Have We Failed to Learn From Past Reforms?

- We have mistakenly believed that teaching is easy.
- We have failed to define the independent variable.
- Research occurs after the fact.
- Educational reform is ahistorical.
- Reform encourages the same for all.
- Newness sells.
No Single Instructional Behavior or Classroom Variable Relates Consistently and Uniquely to Student Achievement: Some Examples

- Clarity
- Teacher Verbal Ability
- Teacher Subject Matter Knowledge
- Teacher Praise
- Assignment of Homework
- Individualizing Instruction
- Classroom Size
- Homework
Thank you!

Any questions?

Other view points?
High-Stakes Teacher Evaluation: High Cost — Big Losses

As schools face public opinion plus state and federal policy changes, principals and teachers are evaluated partly on student performance. The advantages have been widely touted without serious attention to the policy consequences. One criticism: This may push schools to fire teachers based on poor student evaluations. The High-Stakes Teacher Evaluation Conference will present a thoughtful analysis with an eye toward policy issues and possible negative consequences.
Teacher Effects

Spyros Konstantopoulos
Michigan State University
spyros@msu.edu
Importance of Teachers

- Teachers are an important part of the school enterprise.
- Teachers can affect student achievement significantly (anecdotal evidence and empirical research)
- A fundamental objective of teacher effects research is to examine how teachers improve academic achievement.
- NCLB has mandated state plans to improve the effectiveness of teachers. The underlying belief is that effective teachers will increase student achievement.
Reliability of Teacher Effects

- Are indexes of teacher effectiveness based on state tests reliable?

- Can the indexes identify reliably effective and ineffective teachers?

- Some evidence from Florida indicates that reliability estimates vary from 0.30 to 0.80 (McCaffrey, Sass, Lockwood, Mihaly, 2009).

- Lower reliabilities suggest larger sampling error.
Reliability of Teacher Effects

- Indexes of teacher effectiveness are not always reliable.

- The reliability increases when multiple years of data are used (Glazerman et al., 2010).

- Indexes or teacher effectiveness are not less reliable than indexes used for high stakes decisions in other disciplines (Glazerman et al., 2010; McCaffrey et al., 2009)
Stability of Teacher Effects

- The stability or consistency of teacher effects is also important.

- Identify teachers who consistently succeed or struggle in classrooms.

- Does an effective (or ineffective) teacher in one occasion remains similarly effective (or ineffective) in another occasion?
Stability of Teacher Effects

- Elementary school: Different cohorts of students are taught same or similar material by the same teacher from year to year (Rosenshine, 1970).

- Middle-school or high-school: Teachers teach the same lessons in different classes (same day, week, or year).

- The main idea is to examine whether teacher effectiveness remains stable or changes in different settings and time points.
Stability of Teacher Effects

- Early narrative reviews discuss findings from studies that had focused on the stability of teacher effects across instructional periods during a single school year (Rosenshine, 1970; Emmer, Evertson & Brophy, 1979).

- The results indicate relatively low stability of teacher effects from class to class.

- The results are rather similar across time periods. Brophy (1973) found that only some teachers were consistent in their effectiveness over time.
Stability of Teacher Effects

- There is a gap in the literature from the 1970s until very recently (mid 2000s).

- Ballou (2005) found that teachers tend to move between performance rankings (e.g., lowest-quartile to median or highest-quartile to median) relatively frequently over a five-year period.

- Koedel & Betts, (2007) report that teachers in either the top or bottom performing quintiles tend to remain in the respective quintiles over time (unlike teachers closer to the middle of the performance distribution).
Stability of Teacher Effects

- Aaronson, Barrow, & Sander (2007) found a relatively high degree of stability in effectiveness only in the top performance category (deciles).

- The majority of high performing teachers remained in the same category in subsequent years, but most other teachers switched among deciles.
Stability of Teacher Effects

- Measures of teacher effectiveness across different sections and years are correlated in grades 4 through 8 (Gates Foundation).

- However, these correlations are generally modest (i.e., < .50).

- Indexes of stability were higher in mathematics than in reading.
Stability of Teacher Effects

- Similar results were reported by Goldhaber and Hanson (2010). The authors investigated the stability of teacher effects over a 10-year period (grades 3 through 5).

- The results indicated some stability of teacher effects over time. There was greater stability in mathematics than in reading. Also there was greater stability among the highest performing teachers.
Stability of Teacher Effects

- Other recent work has argued that teacher effectiveness ratings often vary considerably from year to year (Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012).

- For nearly 50 percent of the teachers, effectiveness rankings change by at least two deciles from one year to the next.

- The results hold for high (and low) effective teachers as well.
Stability of Teacher Effects

Should education policymakers assume a particular approach to identifying teacher effectiveness will produce a reliable or stable performance?

Are the findings robust and conclusive enough to warrant decisions about raises, tenure, or termination of employment? Can recent findings be used for accountability purposes? Are the criteria that inform the accountability system adequate (are we there yet)? Are we better off now to make such decisions about teachers than 35 years ago? (see Good, Biddle, & Brophy, 1975)
Studies that Examine Teacher Effects

- Studies of associations (e.g., education production function studies).
- Process-product studies.
- Variability in teacher effects studies.
- Value-added models studies.
Teacher Characteristics and Student Achievement

- Earlier reviews have provided inconclusive findings (Hanushek, 1986; Greenwald, Hedges & Lane, 1996).
- However, teacher experience seems to be related to student achievement (Clotfelter, Ladd, & Vidgor, 2006; Nye, Konstantopoulos, & Hedges, 2004).
- Content knowledge also seems to be related to student achievement (Hill, Ball, Rowan, 2005).
- National Board certified teachers are more effective than other teachers (Goldhaber & Anthony, 2007). However, certification is not always related to student achievement.
Process-Product Studies

- Examine association between teaching practices or what teachers do in the classroom and student achievement (e.g., Brophy & Good, 1987; Good & Grouws, 1977; Good, 1979).

- Successful processes include teachers’ confidence in teaching students successfully, efficient allocation of classroom time to instruction and academic tasks, effective classroom organization and group management, and active/engaging teaching that emphasizes understanding of concepts (Good & Brophy, 1987).
Variability in Teacher Effects Studies

- Examine the variance in achievement between teachers controlling for student background.

- In regression models the overall teacher effects are represented by a group of teacher specific binary indicators (not specific characteristics).

- Measure change in $R^2$ (or $\Delta R^2$) between regression models that is due to teacher effects.

- Teacher effects are represented as between-teacher residual variance in student achievement.
Variability in Teacher Effects: Magnitude of Effects

- A considerable amount of variation in student achievement is explained by teacher effects net of student background (e.g., Nye et al., 2004; Rivkin, Hanushek, & Kain, 2005; Rowan, Correnti, & Miller, 2002).

- Nye et al. (2004) reported gain estimates that were nearly 1/3 of a standard deviation (SD) in mathematics and reading in early grades.

- Hanushek and Rivkin (2010) reported gain estimates that ranged between 1/10 and 1/5 SD in mathematics and between 1/10 and 1/3 SD in reading. By way of comparison the effect of one year in small classes on achievement gains in Project STAR was nearly 1/7 SD.
Differences in Teacher Effectiveness

- A common finding in the literature is that teachers differ dramatically in their effectiveness.
- Evidence from earlier work (Brophy, 1986, 1988; Good et al., 1975).
- More recently, Nye et al. (2004) provided similar evidence using data from a large-scale experiment (Project STAR).
- Rivkin et al., (2005) also reproduced this finding using panel data from a large-scale study in Texas.
Value-Added Models

- Determine the unique contribution of teachers on student achievement gains (e.g., McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Meyer, 1997).

- Controlling for student background and prior achievement is essential to eliminate confounding due to non-random assignment of students and teachers to classes.

- Value-added models facilitate causal inferences to a degree, but do not eliminate confounding effects due to unobservables completely (Braun, 2005).
Value-Added Models

- Value-added models are similar to the previous research traditions, since by and large all traditions aim to equate for preexisting differences in student background.

- Recent work has shown that common assumptions in value-added models don’t always hold and that teacher effects estimates can’t be interpreted as causal in some cases (Rothstein, 2010).

- Similarly Rubin, Stuart and Zanutto (2004) argue that causal estimates of teacher effects are difficult to conceptualize and that value-added models do not necessarily produce causal estimates.
Persistence of Teacher Effects

- The persistence of teacher effects on student achievement has been demonstrated (e.g., Ballou, Sanders, & Wright, 2004; McCaffrey et al., 2004; Sanders & Rivers, 1996).

- Sanders and Rivers (1996) found that teacher effects in grades 3, 4, and 5 predicted fifth-grade achievement. Teacher effects were also cumulative.

- Recent work has also demonstrated that teacher effects persist in elementary grades and that their cumulative effects are considerable (Konstantopoulos & Chung, 2011).
Summary

- Findings are mixed and inconclusive:
  - Observed teacher characteristics are not always related to student achievement (Hanushek, 1986)
  - However teachers differ considerably in their effectiveness (Nye et al., 2004; Rivkin et al., 2005)
- Difficult to interpret the relation between teacher characteristics and student achievement as causal (because of selection).
- Can causality be achieved with value-added models? We should be cautious.
Limitations in Teacher Effects Research

- Teacher effects can be confounded with student or family characteristics (student background) because assignment to classes is not random (selection bias is likely).
- Difficult to make inferences about the causal direction of the association between teacher characteristics and student achievement.
- Random assignment of students and teachers to classrooms is more likely to facilitate causal inferences.
- Can we measure and control for all important covariates to eliminate selection?
Teachers Effects for All or Some Groups?

Findings are inconclusive:

- Teacher experience is positively and significantly linked to the achievement of black students (Murnane & Phillips, 1981).
- Teacher characteristics overall don’t seem to be related with the achievement of minority students (Hanushek, 1971, 1992).
- Minority and disadvantaged students benefit as much as other students from effective teachers (Konstantopoulos, 2009; Sanders & Rivers, 1996).
- Differences in teacher effectiveness are more pronounced in low SES schools (mix of more and less effective teachers) (Nye et al., 2004).
Summary of Findings Overall

- There is consistent evidence from experimental or observational data that there are substantial differences among teachers in the ability to produce achievement gains in their students.

- However, teacher characteristics such as education, experience, licensure, or salary have not been consistently related to student achievement.

- Evidence from Project STAR indicates that only 1-2 percent of the differences in teacher effectiveness is due to education and experience.

- Can we identify and measure critical teacher characteristics? Can we measure teacher characteristics or effects (e.g., variability) reliably (free or error)?
Discussion

- We need more evidence about the reliability and the stability of teacher effects.
- Measuring the stability of teacher effects is essential. Use value-added measures of teachers across multiple years.
- It is unclear that the findings thus far warrant accountability related decisions that can seriously affect teachers’ employment, pay, etc.
- We need well thought out and well developed criteria to guide the accountability decisions (Good et al., 1975). Will the criteria be standardized across districts and states?
Discussion

- Agree on parameters that compose the measures used to evaluate teachers. Which get to have more weight?
- Identify and measure factors that explain most of the variation in teacher effectiveness.
- Minimize error and measure “pure” teacher effects.
- Follow-up suggestions from previous work by Good and Brophy: observe systematically teaching practices in classrooms and interview teachers are effective teaching (smaller scale work).
- Design large-scale studies either experiments or well thought out quasi-experiments that aim to eliminate the selection process.
Discussion

- Value-added models are useful, but more work is needed on statistical modeling especially for longitudinal designs.

- Evidence across different states is important to examine whether estimates converge. Are the parameters used to produce indexes of teacher effectiveness similar?

- Identify school context that maximizes teacher effects (e.g., class size, classroom composition, teacher interactions, leadership, etc).
As schools face public opinion plus state and federal policy changes, principals and teachers are evaluated partly on student performance. The advantages have been widely touted without serious attention to the policy consequences. One criticism: This may push schools to fire teachers based on poor student evaluations. The High-Stakes Teacher Evaluation Conference will present a thoughtful analysis with an eye toward policy issues and possible negative consequences.
State and Local Efforts to Investigate the Validity and Reliability of Scores from Teacher Evaluation Systems

High-Stakes Conference
University of Arizona
October 12, 2012
Overview

• Introduction and purpose
• Sample and Methods
• What do we mean by valid and reliable?
• Findings
  ▫ Reliable and Valid
  ▫ Evaluators, Instruments and Sampling Lessons
  ▫ Decisions, PD for Teachers, and Changes in School Culture
Purpose

- We examine states’ efforts to investigate and ensure the reliability and validity of scores emerging from new teacher evaluation systems.
- We focus on areas known to be of concern in the production of high-quality observational scores, including:
  - The choice of the observational instrument,
  - Rater training and certification, and
  - The number of lessons collected per teacher per year.
Sample of 17 States

- Each state received a RTTT grant or NCLB waiver before July 1, 2012;
- Conducted a pilot-test of its new teacher evaluation system during or before the 2012-13 school year;
- Had statutory language describing a teacher evaluation system which satisfied the requirements of RTTT or NCLB waiver;
- And did not have any pending legislation, as of 7/1/12, which would substantially change the system.
Methods

• Document collection and analyses to answer basic questions about state-level guidance
  ▫ Legislation and government guidelines,
  ▫ RTTT applications, and
  ▫ NCLB flexibility requests, etc.
• Led to creation of a 25-question interview protocol
• Interviews with 13 state DOE staff from 12 states, typically a director, coordinator or executive officer of the state’s efforts to implement new teacher evaluation system
What do we mean by reliable and valid?
What do states mean for scores to be reliable and valid?
What efforts are underway to ensure the reliability and validity of scores?
What would you like to see in order to assure reliable and valid scores?
### Making the Grade in New York City

Do the progress reports really help develop good schools? Or do they simply confuse parents and irritate teachers and principals?

<table>
<thead>
<tr>
<th>CATEGORY</th>
<th>SCORE</th>
<th>GRADE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student Progress</td>
<td>18.1</td>
<td>D</td>
</tr>
<tr>
<td>School Environment</td>
<td>7.8</td>
<td>B</td>
</tr>
<tr>
<td>Closing the Achievement Gap</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td>Overall Score</td>
<td>36.9</td>
<td>C</td>
</tr>
</tbody>
</table>
What do we mean by reliable and valid?

- **Score validity**
  - **Construct validity**
    - Face validity
    - Factor analyses or other construct identification procedures
    - Criterion-related validity
  - **Consequential validity**
    - Both for the individuals in the system
    - And for the system itself – are there positive or negative unintended consequences?
What do states mean for scores to be reliable and valid?

- 9 states responded
- The most common answer (3 states) was to see congruence between observation and student-assessment-based metrics
- Two answers focused on elements of the system that would improve validity and reliability
  - a distribution of scores that better reflects reality
  - raters who are more grounded in benchmarks and evidence
- Some answers were very vague, only loosely related to standards for validity and reliability.
What efforts are underway to ensure validity and reliability of the scores?

• 12 states responded
• Responses varied widely across states
• Taken together states are attending to many of the issues involved in establishing reliability and validity of assessment systems
• Rare to see a state attending to more than a handful of issues at a time, and some states are attending to only one.
What efforts are underway to ensure validity and reliability of the scores?

- Enhanced training of evaluators to meet standards
- Empirical investigations (finished or planned)
  - Correlate VAM and observation scores
  - Investigate inter-rater reliability
  - Factor analyses of data
  - Studies of validity and reliability studies
- Auditing/monitoring techniques planned
- Selection of instruments with evidence of reliability and validity
What would you like to see in order to assure reliable and valid scores?

- 8 states responded
- Responses widely varied across states, and, where we interviewed more than one official per state, variable within states as well.
- Two states were concerned with technical aspects of the production of VAM scores
- Two states expressed concern about the reliability and validity of assessments in non-tested subjects
- Two states suggested a process for improving reliability and validity
- Other state concerns involved widening the pool of evaluators
Evaluators, Instruments and Sampling

Who is required to conduct teacher observations?
How did states go about choosing the instrument?
How did states arrive at the number of required observations?
Are there plans to report reliability and validity of scores?
Questions about the evaluators

- Have you thought about how to account for possible subjectivity from raters? For example, principals/ coworkers assigning overly harsh or lenient scores?
- Is your state/office involved in the training of raters of classroom instruction? How?
- Is your state/office involved in the certification of raters of instruction?
- Is your state/office involved in the monitoring of raters once they pass certification?
- Are there checks in place to ensure validity and reliability of the scores coming in? For example, will there ever be a second scorer to cross check scores assigned by raters? Or will you ever look at trends of overall scores from principals?
- Do raters periodically have to recertify?
- Are you planning on monitoring/regulating inter-rater reliability in the field? What are your basic plans for the collection, storage, and use of the data from the observations? Are there data systems in place to record, trace, or flag data? How will you flag outlying or extreme examples?
What we learned about evaluators

- 17 states responded
- Principals are primary and lone evaluators
- All evaluators will participate in some type of training
- Training usually focuses on familiarizing raters with the evaluation process, rubrics and types of evidence, although some training also includes how to provide effective feedback
Evaluators

Who is required to conduct teacher observations?
Principal/Administrator: (DE, FL, LA, MD, TN, NY, IL)
Principal/Administrator OR instructional leader, other designee: (CO, IN, MA, OH, RI)
Principal/Administrator and peer (NC)
No specification (AZ, CT, GA, OK)

Do states require multiple raters?
Yes (NC)
Recommend (IN)
Conditional, only for an evaluation of ineffectual (MD)
No (AZ, CO, CT, DE, FL, GA, IL, LA, MA, NY, OH, OK, RI, TN)
What we learned about evaluators

- Rater training is both centralized and un-centralized, and in-house and by contract firms.
- Although some states talk about certification, very few describe an object evaluation process for raters.
- Many states will likely award certification for attending training sessions.
- Very few states specify a timeline for requiring rater recertification.
- Many states plan to provide ongoing professional development and support to raters but this is rarely an objective monitoring process.
Who is responsible for training raters?
State training (DE, GA, IL, MD, MA, OH, TN)
District training (CO, CT, FL, IN, LA, NY)
No specification or Other (AZ, NC, OK, RI)

Do states require raters to meet objective certification criteria (other than simply attending a training workshop)?
Yes (IL, LA, OH, TN, NC, OK)
No (AZ, CO, DE, GA, CT, FL, IN, MA, MD, NY, RI)
Questions about observational instruments

- How did you go about choosing the instrument?
- What did you know about the tool before you adopted it?
  - Did you have any information about reliability of validity?
- What do you know about the way it has been adapted by districts?
What we learned about instruments

• 11 states responded to question about how the instrument(s) was chosen
  ▫ Most instruments were chosen by a committee or task force to align with state standards
  ▫ Five states created their own tool; others uses an existing/modified tool(s)

• 9 states responded to questions about how the instrument has been adapted by districts
  ▫ Unless the state model was mandated, few states knew of district implementation plans
Type of Evaluation System

- Entirely state-based system (DE, GA)
- State model(s) with some district choice (IL, MD, MA, NC, OH, OK, RI, TN)
- Mainly district developed (AZ, CO, CT, FL, IN, LA, NY)
What we learned about instruments

- 10 states responded
- Most states reviewed some information about the validity and reliability of the instruments chosen
  - Two states did their own study
  - One state plans to do its own study
  - Three states looked at data provided by outside sources
  - Four states did not know about the reliability or validity of the instrument
Questions about sampling of lessons

• We know that statutes require X observations per teacher per year. Can you tell me a little about how you arrived at that number?
• Was any kind of study conducted using your instrument in order to determine the optimal / most cost efficient number of observations? If so, when? If not, why not?
• Is there any flexibility around this number? Off-the-books guidance to districts around the number of observations?
• Have you determined if any of these observations will be unannounced? Why or why not?
What we learned about sampling of lessons

- 10 states provided information about how the number of observations was determined.
- None indicated that there was a formal study to determine the optimal number of observations.
- Four states mentioned using research from other studies (MET was mentioned three times) to determine the number.
- Five states indicated that the number was arrived at through a negotiation process with teachers' unions.
- Five of the states indicated that the number was chosen, at least in part for practical and logistical reasons.
- Two states said that raters are encouraged to conduct more observations if needed.
States by Number of Observations Required

- 0 (OK)
- 1 (FL, DE)
- 2 (AZ, CO, GA, IL, IN, LA, MD, NY, OH)
- 3 (CT, MA, NC, RI)
- 4 (TN)
Are there plans to report reliability and validity of scores?

- 11 states responded
- Few states have plans to publish reliabilities, though many states plan to conduct research on the new systems
- Only 3 states (of 17 in full sample) have legislation that requires a report on aspects of reliability and validity
State specifies inquiry into overall scores

Inquiry into value added scores

Inquiry into observation system
Decisions, PD for teachers, change in school culture

What purposes will the overall teacher evaluation scores be put?
Are there efforts to ensure those decisions are correctly made?
What kinds of PD are teachers receiving about the new systems?
Do you think the evaluation system may change the culture in schools?
What purposes will the overall teacher evaluation scores be put?

- 17 states responded
- All but one state (AZ) specify consequences of poor performance on a teacher evaluation, ranging from not receiving raises, participating in assistance programs and remediation plans to dismissal and termination.
- All but one state (CT) specify some consequence for excellent performance, ranging from bonuses, salary increases, non-probationary status, tenure, and fewer observations in the future.
Are there efforts to ensure those decisions are correctly made?

- Only 3 states responded
  - Few states have specified ways to ensure correct decision making.
- Most states say they are waiting to see how the new system works and may make adjustments
What kinds of PD are teachers receiving about the new systems?

- 12 states responded
- 10 indicated that professional development and information for teachers was largely left to districts
- One state had been directly involved in informing teachers by delivering presentations and talks to different schools
- One state that had recently passed legislation charging the state to provide training on the evaluation system to all teachers
What kinds of PD are teachers receiving about the new systems?

- There were four states that were using variations of "train the trainer" models.
- Two states described providing building level principals "meetings in a box" to disseminate information about the evaluation systems.
- Four states mentioned that they were providing information about the evaluation and training opportunities through online platforms. Webinars were mentioned by officials from two states.
Do you think the evaluation system may change the culture in schools?

• 7 states responded
• All respondents acknowledged to some degree that the new evaluation systems would identify fewer teachers in the top tier than the previous system did
• Four of the states specifically mention that the transition will be difficult
• Four states talk about the difficulties in redefining the ratings of "satisfactory" and "proficient" so that the expectation isn't that all teachers should pile up at "exemplary" category
Do you think the evaluation system may change the culture in schools?

- In one state, there are plans for principals themselves to be evaluated on their ability to be good evaluators, which may, the official said, help to temper any perverse incentives principals could have to assign inflated scores.
- Finally, one state framed the shift to identifying fewer exemplary teachers as quite positive, saying,
  - "It’s going to come down to embracing the principles behind the model and being able to communicate and message that throughout schools and districts while really fostering a sense of trust, that this is not about - it’s not a gotcha, but this is about identifying, recognizing truly excellent practice, ...seeing this as a collective effort to get better... knowledge is power."
Conclusions

• An important goal for research and policymakers is to investigate whether the right decisions are made based on evaluation scores
• Few seem to be considering the negative unintended consequences of a system that may generally be perceived of as being arbitrary
• There has been huge change in a very short time, and there may be good reasons why states are falling short in the effort to balance better systems and available resources—but, states are not necessarily making informed trade-offs
If you have questions or would like more information, please contact:

Corinne Herlihy  
Project Director  
National Center for Teacher Effectiveness  
Harvard Graduate School of Education  
50 Church Street, 4th Floor  
Cambridge, MA 02138  
617-496-9817  
corinne_herlihy@gse.harvard.edu
High-Stakes Teacher Evaluation: High Cost — Big Losses

As schools face public opinion plus state and federal policy changes, principals and teachers are evaluated partly on student performance. The advantages have been widely touted without serious attention to the policy consequences. One criticism: This may push schools to fire teachers based on poor student evaluations. The High-Stakes Teacher Evaluation Conference will present a thoughtful analysis with an eye toward policy issues and possible negative consequences.
Caught in a Vise:
The Challenges Facing Teacher Preparation in an Era of Accountability

Rick Ginsberg, Ph.D. & Neal Kingston, Ph.D.
University of Kansas
The Causality Link

For those NOT involved in education, this sort of "value-added" approach to education makes perfect sense, but to those of us who are intimately involved in this calling understand that boiling education down to a single number is an exercise in futility. Should we all be held accountable? Absolutely! After all, we are talking about our most precious commodity - our children and their intellectual and emotional growth and well-being. But have we figured out the ultimate best method for this accountability system yet? Absolutely not!

- a State Department of Education leader, former superintendent, principal and teacher (Personal communication, September 19th, 2012)
The Causality Link

The Torcetrapid Story

Correlation does not imply causation!!
The Causality Link
The Causality Link

• Explanations:
  – Increases in population
  – Seat Belt Use
  – Stricter Speed Limits
  – Media focus on traffic safety
  – Crack down on drunk driving
  – Safer cars
  – NAFTA
  – Weaker crop output in U.S.

• Lemons may prevent scurvy and have multiple health benefits, but this relationship....!!

• *Cum hoc ergo propter hoc* (with this, therefore because of this) --It is the art of the **logical fallacy** – when we assign causality to correlated variables
Our Research

• Explaining the Vise - the current context for teacher preparation accountability
• The use of outcome measures in a number of professions
• Strengths and weaknesses of emerging assessment models and measures
• Potential avenues to explore
Current Context

• What is the vise teacher preparation finds itself in?
Constant Criticism of Teacher Preparation

• A Nation at Risk (1983)
• NCTQ
  – States on web-site: “Unlike other professional schools, teacher preparation programs are held to weak standards, enabling ineffective programs to receive state approval and national accreditation. The result? Too few teachers receive the knowledge and skills they need to be successful in the classroom.”

• Levine Study (2006)
• Popular Media: George Will (2006)
  – “the surest and quickest way to add quality to primary and secondary education would be addition by subtraction – close all the schools of education!”
Constant Criticism - II

- Secretary of Education Arne Duncan (speech at TC, 2009):

  “By any standard, many if not most of the nation’s 1,450 schools, colleges, and departments of education are doing a mediocre job of preparing teachers for the realities of the 21st century classroom.....”
Confounding Findings

• 2008 Public Agenda survey of teachers
  – 80% of teachers reported they were very (42%) or somewhat (38%) prepared for their first year of teaching

• Gallup poll results
  – Public suggests no sense of crisis:
    • 71% have trust and confidence in public school teachers
    • 77% give the school their eldest attends an “A” or “B”
    • Describe characteristics of teachers who made a difference in their lives (“caring,” “encouraging,” “attentive/believe in me”)
Confounding Findings - II

• 1st CAEP Annual Report (2012)
  – NCATE results, 2011:
    • 135 accreditation decisions, 67% fully accredited for 7 years (many with some AFIs), 9% approved after having to go through a review two years earlier, 13% approved for only two years with a need for additional site visit, 8% deferred for needing more information, 3% had accreditation revoked, 3% dropped out pending a negative decision

• AACTE PEDs Results (2012)
  – Clock Hours (BA programs) - early field experiences + student teaching:
    • Low average programs = 614 hours
    • High average programs = 751 hours
The Vise

- Constant criticism
- Ample evidence doesn’t jive with much of the criticism
- Educators agree with the need for improvements in terms of assessing student performance
- Squeeze from demands for accountability (RTTT, Higher Education Act – Neg Reg) based on measures that are error prone and highly questionable in terms of capturing their intent (question of validity) – Back to the causality link
Professional Accreditation

• Push for measuring student outcomes is present in all professions
• CHEA highlights direct and indirect measures to consider:
  Direct: capstone performances, professional/clinical performances, third party testing (e.g. licensure), and faculty designated examination
  Indirect: portfolios and work samples, follow-up of graduates, employer rating of graduates, and self-reported growth by graduates
Professional Accreditation II

- ABA Outcome Measures Committee (2008)
  - Studied 10 professions
  - Identified 28 assessment criteria used across all professions
    - Range of criteria used – 5 to 16
      - 5 – Dentistry, Engineering
      - 15 – Medicine, Pharmacy
      - 16 – Teacher Education
  - Six outcome measures most common:
    - licensure of graduates – 7
    - clinical, problem solving and communication skills of students – 7
    - criteria to insure that students possess the competencies expected by the profession and public – 6
    - evaluation of the skills, knowledge and behavior/attitudes of students – 6
    - student portfolios – 5
    - collection of evidence-based data of learning objective competencies – 5)
  - Teacher Education the only profession that does all six!
Accreditation III

• Medicine as a case example – linking training to patient care is challenging:
  – Liaison Committee on Medical Education (LCME) - MDs
  – Accreditation Council for Graduate Medical Education (ACGME) - Residency sites for MDs
  – Concern with “Quality of Care”
    • (Swing, 2007) – “measurement of patient care quality, in particular using clinical process and outcome measures, is still in its infancy”
    • Haan, et al. (2008) – “Medical education does not occur in isolation, and most process and outcome measures represent the group milieu in which treating and learning occur”
    • Rand Europe (2011) – “available evidence of direct association between the quality of healthcare education and training and the quality of care provided remains scant.”
Professional Accreditation IV

- We studied 10 professions (Law, Medicine, Psychology, Journalism, Pharmacy, Social Work, Engineering, Athletic Training, Business, Teaching)
- Identified accrediting agencies, examined accreditation standards and related documents, spoke with accreditation directors
- Conclusions:

  1) Programs typically have a great deal of leeway in devising assessment systems and plans.

  2) Accrediting bodies differentiate between direct (proximal, internal) vs. indirect (distal or external) measures. Greater reliance and preference for direct measures.

  3) Much institutional isomorphism among the professions – a lot of similarities.

  4) After graduation work performance as an indicator of pre-service training success – if used at all – largely relies on some kind of employer satisfaction or graduate/alumni self-efficacy surveys.
Professional Accreditation V

• Measuring outcomes in the workplace:

  – “We use alumni surveys which are marginally effective. We attempt employer surveys which are even more problematic....Advisory Boards also provide a source since they hire our graduates.” (Engineering)

  – “We talk about other potential measures including patient safety/care or more specifically quality of patient care...and health care cost savings...However, it is a long, long way to relating these types of things to the education an individual receives at a particular institution.” (Pharmacy)
Exploring the Causality Link – Numbers, Valued-Added, Growth and Other Random Thoughts

• Numbers Don’t Lie (but they do keep their fingers crossed!!)
  – The Prediction Machine – mathematically models, analyzes and handicaps all sports
  – College Football home field advantage by team
    • University of Kansas - #14
    • Remarkable – KU’s home field record from 2009-2011 was 8-13 (did lose more at away games because schedule cupcakes at home early in season)
  • How?
    – Formula looked at swings in performance between home and away games (indeed, in same period KS was 2-13 away)
Number Games

• Just to make the point, in basketball, here are the rankings for home court advantage for last year’s final four teams
  – Ohio State - #47
  – Louisville - #122
  – Kansas – #86 (with 59 straight home wins)
  – Kentucky - #134 (national champion)
Number Games II

• **It isn’t just sports!**

• Internet Legal Research Group (ranks law schools by bar passage rates):
  – #1 - (tied) Marquette
  – #9 - Campbell University
  – #10 – Ave Maria School of Law
  – #36 – Yale University (#1 in U.S. News and World Reports – lowest acceptance rate in U.S.)
  – #64 – Stanford (#2 in U.S. News and World Reports - second lowest acceptance rate in U.S.)
Concerns With Relying on Tests

- Iatrogenic diseases
- Negative externalities
- Campbell’s Law
  - "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor."
Concerns With Relying on Test Scores

• Curriculum and Instruction
  – Teaching to the test
  – Shrinking the curriculum

• Students
  – Non-random distribution of students to schools and classrooms

• Teachers
  – Non-random distribution of teachers to undergraduate and graduate programs
  – Students take courses, credentials and degrees from multiple institutions
  – Doesn’t work for all teachers or educators
  – Quells teaching creativity (Zhao, 2012)
Concerns With Relying on Test Scores II

• Tests:
  – Lack of reliability of difference scores (Time 2 – Time 1)
  – Decision Consistency (e.g. MET study review: “teacher evaluations based on observed state test outcomes are only slightly better than coin tosses at identifying teachers whose students perform unusually well or badly on assessments of conceptual understanding.”
  – Regression to the mean
  – Tests capture a limited perspective of student learning
  – Causal claims may not be warranted (CCSSO Study, 2012)
  – Instructional insensitivity
More on Instructional Insensitivity

• If you expect better teachers to increase test scores then wouldn’t you expect students to do better on items where the topic was taught versus those where the topic was not?
  – 95% of TIMMS items were shown to be instructionally insensitive

  • Much learning goes on outside of schools – how can teachers be held accountable for that?
  • The statistical methods used to select test items may cause an over-selection of items that load heavily on general ability rather than achievement
Value Added Means different things

- Student growth percentiles
- One step VAMs
- Two step VAMs
  - Issues (Braun, 2008):
    - Small sample volatility
    - Properties of vertical scaling
    - Missing data
    - Estimates based on VAMs are descriptions
  - Concludes: “Teachers bear a substantial degree of responsibility for the learning of their students, but the degree varies with the school and community, the mix of students in the classroom, the number of ‘unexpected events,’ etc.” OR how do we measure true teacher effectiveness?
It’s Not the Tests – It’s How They’re Used
So We’re Back to Torcetapid

• Ignoring the causality link will lead to inaccurate judgments
• Tests offer descriptions
• Causation is a set of observations, not independent facts
• Need to take care in judging individuals based on tests alone
• Even in a multiple measure weighted system, requiring some level of test performance makes that indicator dominant (the quantified metric drives the decision - Baker, 2012)
Releasing the Vise

• Get the word out: We’re not as bad as they say we are!!
• Regarding accountability for teacher preparation focusing on student learning:
  – Multiple measures, multiple times, over multiple years (Ron Ferguson)
  – edTPA (or other portfolio-types of approaches)
  – Tripod Project – surveys focusing on student engagement and classroom learning conditions, the 7 C’s (care, control, clarify, challenge, captivate, confer, consolidate)
  – Observations
  – Create a version of “Habits of the Mind” (Meier)
  – Use tests as a component, an indicator, not as an end in and of themselves
High-Stakes Teacher Evaluation: High Cost — Big Losses

As schools face public opinion plus state and federal policy changes, principals and teachers are evaluated partly on student performance. The advantages have been widely touted without serious attention to the policy consequences. One criticism: This may push schools to fire teachers based on poor student evaluations. The High-Stakes Teacher Evaluation Conference will present a thoughtful analysis with an eye toward policy issues and possible negative consequences.
Exogenous Variables and Value-Added Assessments: A Fatal Flaw if Value-Added Assessments are Linked to High-Stakes Decisions

David C. Berliner
Regents’ Professor Emeritus
Mary Lou Fulton Teachers College
Arizona State University

Tucson AZ, University of Arizona, October, 2012
The politicians, business leaders, and the public currently think of the influence of a teacher on a student like this:
But everyone who has taught knows that the relationship is much more likely to be reciprocal, more like this:

\[ T \leftrightarrow S \]

{In fact, all those politicians, business leaders, and anyone among the general public who have been parents or ever had a pet, knows full well that their simple model of influence is ridiculous, since relations with every child they have and all pets they ever had depends on characteristics of the child or the pet. Influence goes both ways! Why people cling to stupid beliefs is another story!}
And if it we think about classrooms, not a single student, then the relationship of a teacher to the students is more complex and looks more like this:

![Diagram of teacher and students relationship]

But students also have relationships with each other. They form small clusters (the nerds, the jocks, the cheerleaders, the screw-ups, the gifted, the special education kids, etc.), and thus these 25 or more reciprocal individual relationships that a teacher maintains, each of which affects achievement, are also supplemented by dozens more reciprocal relationships with the students as members of their groups, and that is plural, because a student has membership in multiple groups. The number and power of, say, the “compliers” and the “resisters” or the “anxious” and the “non anxious” in a class can dramatically affect the outcomes of schooling in a particular year. Yet such characteristics of classrooms are unmeasured and may actually be unmeasurable, a major problem for value-added approaches to assess teachers because it is a given in all value-added approaches that sources of influence on student achievement scores are knowable and/or controllable. But this problem is merely the tip of the exogenous variable iceberg. The real issues are the outside-of-classroom variables affecting achievement inside the classroom.
**Exogeny**: In an economic model an exogenous change is one that comes from outside the model and is unexplained by the model.

So here is the real state of affairs: Each $\Rightarrow$ = one or more variables that are unexplained by any value added model of teacher effectiveness. Such variables are endless.............

**Results**

Correlations in **Mathematics** between a large set of teacher behavior variables and value-added scores:

- Across years: .12 to .25
- Across different sections of the same subject in the same year: .16 to .26

Correlations in **English/Language Arts** between a large set of teacher behavior variables and value-added scores:

- Across years: .09 to .12
- Across different sections of the same subject the same year: .10 to .24

Millions of dollars spent to find out that **teacher behavior** is not stable across raters, across occasions, and across instruments purporting to assess the same behavior, and also that **teacher effects on student achievement** are not stable across years and across section of the same class in the same year. The researchers are still scratching their heads!
<table>
<thead>
<tr>
<th>Changes in teachers rankings across two courses taught in the same year</th>
<th>The change in teacher ranking in effectiveness was one or more Deciles</th>
<th>The change in teacher ranking in effectiveness was two or more Deciles</th>
<th>The change in teacher ranking in effectiveness was three or more Deciles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>85%-%</td>
<td>54%-%</td>
<td>39%-%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
<td>92%</td>
<td>54%</td>
</tr>
<tr>
<td>Changes in teachers rankings across two consecutive years teaching the same type of course</td>
<td>74%-%</td>
<td>45%-%</td>
<td>19%-%</td>
</tr>
<tr>
<td></td>
<td>93%</td>
<td>63%</td>
<td>41%</td>
</tr>
</tbody>
</table>
Year to year value added score in NYC. 0 = no gain compared to students like yours; 99 is maximum gain for students like your own. Year to year stability is not good.

\[ r = 0.35 \]  

About what Brophy found without covariates 40 years ago.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Math</td>
<td>- 2.03</td>
<td>+ .68</td>
<td>+ .16</td>
<td>+ 3.46</td>
</tr>
</tbody>
</table>

**EVASS VALUE-ADDED SCORES**

**GREEN** = SCORES BETTER THAN TEACHERS WITH STUDENTS LIKE THEIRS

**RED** = SCORES WORSE THAN TEACHERS WITH STUDENTS LIKE THEIRS
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Math</td>
<td>-2.03</td>
<td>+0.68</td>
<td>+0.16</td>
<td>+3.46</td>
</tr>
<tr>
<td>Reading</td>
<td>-1.15</td>
<td>-0.96</td>
<td>+2.03</td>
<td>+1.81</td>
</tr>
</tbody>
</table>

**EVASS VALUE-ADDED SCORES**

**GREEN** = SCORES BETTER THAN TEACHERS WITH STUDENTS LIKE THEIRS

**RED** = SCORES WORSE THAN TEACHERS WITH STUDENTS LIKE THEIRS
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Math</td>
<td>-2.03</td>
<td>+.68</td>
<td>+.16</td>
<td>+3.46</td>
</tr>
<tr>
<td>Reading</td>
<td>-1.15</td>
<td>-.96</td>
<td>+2.03</td>
<td>+1.81</td>
</tr>
<tr>
<td>Language Arts</td>
<td>+1.12</td>
<td>-.49</td>
<td>-1.77</td>
<td>-.20</td>
</tr>
</tbody>
</table>

**EVASS VALUE-ADDED SCORES**

**GREEN =** SCORES BETTER THAN TEACHERS WITH STUDENTS LIKE THEIRS

**RED =** SCORES WORSE THAN TEACHERS WITH STUDENTS LIKE THEIRS
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Grade 5</td>
<td>Grade 4</td>
<td>Grade 3</td>
<td>Grade 3</td>
</tr>
<tr>
<td>Math</td>
<td>-2.03</td>
<td>+0.68</td>
<td>+0.16</td>
<td>+3.46</td>
</tr>
<tr>
<td>Reading</td>
<td>-1.15</td>
<td>-0.96</td>
<td>+2.03</td>
<td>+1.81</td>
</tr>
<tr>
<td>Language Arts</td>
<td>+1.12</td>
<td>-0.49</td>
<td>-1.77</td>
<td>-0.20</td>
</tr>
<tr>
<td>Science</td>
<td>+2.37</td>
<td>-3.45</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Social Studies</td>
<td>+0.91</td>
<td>-2.39</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Evass Value-Added Scores**

Green = Scores better than teachers with students like theirs

Red = Scores worse than teachers with students like theirs
<table>
<thead>
<tr>
<th>Author</th>
<th>Country</th>
<th>Some Effects of Peers on Achievement in the classroom, easily confounded with teacher effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>McEwan (2003)</td>
<td>Chile</td>
<td>Mothers education has big effects, fathers education and indigenous students have lesser but still sig. effects</td>
</tr>
<tr>
<td>Author</td>
<td>Country</td>
<td>Some Effects of Peers on Achievement in the classroom, easily confounded with teacher effects</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-----------</td>
<td>-------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>McEwan (2003)</td>
<td>Chile</td>
<td>Mothers education has big effects, fathers education and indigenous students have lesser but still sig. effects</td>
</tr>
<tr>
<td>Lavy, Silva, and Weinhardt</td>
<td>England</td>
<td>Low achieving students affect every ones scores. Boys and girls affected differentially, but both significantly affected by composition of classrooms</td>
</tr>
<tr>
<td>Author</td>
<td>Country</td>
<td>Some Effects of Peers on Achievement in the classroom, easily confounded with teacher effects</td>
</tr>
<tr>
<td>-------------------------</td>
<td>----------------------</td>
<td>------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>McEwan (2003)</td>
<td>Chile</td>
<td>Mothers education has big effects, fathers education and indigenous students have lesser but still sig. effects</td>
</tr>
<tr>
<td>Lavy, Silva, and Weinhardt (2009)</td>
<td>England</td>
<td>Low achieving students affect every ones scores. Boys and girls affected differentially, but both significantly affected by composition of classrooms</td>
</tr>
<tr>
<td>Ammermueller and Pischke (2009)</td>
<td>Six European Nations</td>
<td>Social class composition affects teacher effectiveness</td>
</tr>
<tr>
<td>Author</td>
<td>Country</td>
<td>Some Effects of Peers on Achievement in the classroom, easily confounded with teacher effects</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>------------------------</td>
<td>-----------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>McEwan (2003)</td>
<td>Chile</td>
<td>Mothers education has big effects, fathers education and indigenous students have lesser but still sig. effects</td>
</tr>
<tr>
<td>Lavy, Silva, and Weinhardt (2009)</td>
<td>England</td>
<td>Low achieving students affect every ones scores. Boys and girls affected differentially, but both significantly affected by composition of classrooms</td>
</tr>
<tr>
<td>Ammermueller and Pischke (2009)</td>
<td>Six European Nations</td>
<td>Social class composition affects teacher effectiveness</td>
</tr>
<tr>
<td>Schneeweiss and Winter-Ebmeir (2005)</td>
<td>Denmark</td>
<td>Social/cultural levels of classes affects teacher effectiveness</td>
</tr>
<tr>
<td>Author</td>
<td>Country</td>
<td>Some Effects of Peers on Achievement in the classroom, easily confounded with teacher effects</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>---------------------</td>
<td>------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>McEwan (2003)</td>
<td>Chile</td>
<td>Mothers education has big effects, fathers education and indigenous students have lesser but still sig. effects</td>
</tr>
<tr>
<td>Lavy, Silva, and Weinhardt (2009)</td>
<td>England</td>
<td>Low achieving students affect every ones scores. Boys and girls differentially but significantly affected by composition of classes</td>
</tr>
<tr>
<td>Ammermueller and Pischke (2009)</td>
<td>Six European Nations</td>
<td>Social class composition affects teacher effectiveness</td>
</tr>
<tr>
<td>Schneeweiss and Winter-Ebmeir (2005)</td>
<td>Denmark</td>
<td>Social/cultural levels of classes affects teacher effectiveness</td>
</tr>
<tr>
<td>Schindler-Angvid (2003)</td>
<td>Denmark</td>
<td>Low ability students helped by presence of high ability students, high ability students not hurt by presence of low ability students</td>
</tr>
<tr>
<td>Levin (2001)</td>
<td>The Netherlands</td>
<td>Same as above</td>
</tr>
<tr>
<td>Oakes (2005)</td>
<td>USA</td>
<td>Same as above</td>
</tr>
<tr>
<td>Author</td>
<td>Country</td>
<td>Some Effects of Peers on Achievement in the classroom, easily confounded with teacher effects</td>
</tr>
<tr>
<td>-------------</td>
<td>---------</td>
<td>------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Hoxby (2000)</td>
<td>USA</td>
<td>The % female, % black, % Asian and other compositional configurations affects every one's scores</td>
</tr>
<tr>
<td>Author</td>
<td>Country</td>
<td>Some Effects of Peers on Achievement in the classroom, easily confounded with teacher effects</td>
</tr>
<tr>
<td>-------------------------</td>
<td>--------------------------</td>
<td>---------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Hoxby (2000)</td>
<td>USA</td>
<td>The % female, % black, % Asian and other compositional configurations affects every one’s scores</td>
</tr>
<tr>
<td>Perry (2007) and OECD (2004)</td>
<td>Australia and PISA countries</td>
<td>School social and cultural compositional factors outweighed student’s own social and cultural factors</td>
</tr>
<tr>
<td>Author</td>
<td>Country</td>
<td>Some Effects of Peers on Achievement in the classroom, easily confounded with teacher effects</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-----------------</td>
<td>---------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Hoxby (2000)</td>
<td>USA</td>
<td>The % female, % black, % Asian and other compositional configurations affects every ones scores</td>
</tr>
<tr>
<td>Perry (2007) and OECD (2004)</td>
<td>Australia and PISA countries</td>
<td>School social and cultural compositional factors outweighed student’s own social and cultural factors</td>
</tr>
<tr>
<td>Rumberger and Palardy (2005)</td>
<td>USA</td>
<td>School social and cultural compositional factors equal to student’s own social and cultural factors</td>
</tr>
<tr>
<td>Author</td>
<td>Country</td>
<td>Some Effects of Peers on Achievement in the classroom, easily confounded with teacher effects</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>------------------------</td>
<td>---------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Hoxby (2000)</td>
<td>USA</td>
<td>The % female, % black, % Asian and other compositional configurations affects every one's scores</td>
</tr>
<tr>
<td>Perry (2007) and OECD (2004)</td>
<td>Australia and PISA countries</td>
<td>School social and cultural compositional factors outweighed student’s own social and cultural factors</td>
</tr>
<tr>
<td>Rumberger and Palardy (2005)</td>
<td>USA</td>
<td>School social and cultural compositional factors equal to student’s own social and cultural factors</td>
</tr>
<tr>
<td>Henry and Rickman (2007)</td>
<td>USA</td>
<td>“The ability level of the peers in a child’s classroom has direct...effects on the child’s cognitive skills, pre-reading skills and expressive language skills after controlling for preschool resources, family characteristics, and the child’s skills at the beginning of preschool.”</td>
</tr>
</tbody>
</table>
High-Stakes Teacher Evaluation: High Cost — Big Losses

As schools face public opinion plus state and federal policy changes, principals and teachers are evaluated partly on student performance. The advantages have been widely touted without serious attention to the policy consequences. One criticism: This may push schools to fire teachers based on poor student evaluations. The High-Stakes Teacher Evaluation Conference will present a thoughtful analysis with an eye toward policy issues and possible negative consequences.
Exploring the Implications of High-Stakes Teacher Evaluation on Schools, Teachers, and Students

ALYSON LEAH LAVIGNE, PH.D.
ROOSEVELT UNIVERSITY
Accountability: Modern Day

- Schools, teachers, and students are being held accountable on the basis of high-stakes achievement tests
- “Stakes” are getting higher for teachers
- Criticism of American education continues
- The tension around fair evaluations continues to build
Evaluating Unintended Consequences

- Media discourse has focused on the implementation of teacher evaluations.
- Little attention has been given to the unintended consequences of using teacher evaluations for human capital decisions.
  - In particular, what are the direct and indirect consequences of such decisions for schools, teachers, and students?
To what extent are data from teacher evaluations being used?
- Who has access to teacher evaluation data?
- What are the consequences for a low evaluation?
- Are teachers rewarded for high evaluations?
- Are teacher evaluations tied to certification?
- Are different policies in place for non-tenured and tenured teachers?
How Are Results Used?

- In 2011, 20 states had policies that required teachers to be eligible for dismissal based on evaluation results (NCTQ, 2011)
- According to Race to the Top (RttT), states should:
  (iv) use evaluations, at a minimum, to inform decisions regarding---
  (a) Developing teachers and principals, including by providing relevant coaching, induction support, and/or professional development;
  (b) Compensating, promoting, and retaining teachers and principals, including by providing opportunities for highly effective teachers and principals...to obtain additional compensation and be given additional responsibilities;
  (c) Whether to grant tenure and/or full certification (where applicable) to teachers and principals using rigorous standards and streamlined, transparent, and fair procedures; and
  (d) Removing ineffective tenured and untenured teachers and principals after they have had ample opportunities to improve, and ensuring that such decisions are made using rigorous standards and streamlined, transparent, and fair procedures. (U.S. Department of Education, 2010, p. 19504)
## Consequences Related to Teacher Effectiveness Data

<table>
<thead>
<tr>
<th>State</th>
<th>Tenure</th>
<th>Dismissal</th>
<th>Certification</th>
<th>Compensation</th>
<th>Layoff Decisions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delaware</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D.C.</td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Florida</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Maryland</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New York</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ohio</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rhode Island</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Tennessee</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>6</td>
<td>6</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

(NCTQ, 2011)
**RttT Profile: Tennessee**

- **Teacher Evaluation**
  - 50% student achievement data; 50% observation data, personal conferences, review of prior evaluations and work
  - Teachers are rated on five performance levels:
    - Significantly above expectations, above expectations, at expectations, below expectations, significantly below expectations

- **Use of Evaluation Data**
  - Teachers are eligible for tenure after 5 years and if they have received a rating in the two top evaluation categories
  - Teachers who are tenured post-July 2011, may be subject to having their tenure status removed if they receive two consecutive ratings of 1 or 2 until they receive two consecutive ratings of 4 or 5 ([team-tn.org](team-tn.org))
## Tennessee Results: 2011-2012

<table>
<thead>
<tr>
<th>Level</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>TVAAS Individual Teacher Effect</td>
<td>16.5%</td>
<td>8.1%</td>
<td>24.5%</td>
<td>11.9%</td>
<td>39.1%</td>
</tr>
<tr>
<td>Observation</td>
<td>.2%</td>
<td>2.2%</td>
<td>21.5%</td>
<td>53.0%</td>
<td>23.2%</td>
</tr>
</tbody>
</table>

(Tennessee Department of Education, 2012)
RttT Profile: District of Columbia

- **Teacher Evaluation**
  - 50% student achievement data; 40% classroom observations; 10% commitment to the school community (DCPS, 2012)
  - Teachers are rated on four performance levels:
    - Highly effective, effective, minimally effective, ineffective

- **Use of Evaluation Data**
  - Teachers rated ineffective are subject to dismissal
  - Teachers rated minimally effective are encouraged to seek support and assistance through professional development
  - Teachers rated minimally effective for two consecutive years are eligible for dismissal (http://osse.dc.gov/service/race-top)
## D.C. Results: Two Years of Implementation

<table>
<thead>
<tr>
<th></th>
<th>2009-2010</th>
<th>2010-2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highly Effective</td>
<td>1499 (23%)</td>
<td>1213 (18%)</td>
</tr>
<tr>
<td>Effective</td>
<td>4086 (62%)</td>
<td>4269 (65%)</td>
</tr>
<tr>
<td>Minimally Effective</td>
<td>727 (11%)</td>
<td>750 (11%)</td>
</tr>
<tr>
<td>Ineffective</td>
<td>135 (2%)</td>
<td>113 (2%)</td>
</tr>
<tr>
<td>Ineligible to Score</td>
<td>143 (2%)</td>
<td>238 (4%)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>6590</strong></td>
<td><strong>6583</strong></td>
</tr>
</tbody>
</table>
Real Consequences for Teachers

- **Dismissal**
  - In many RttT states, dismissal policies are tied to teacher evaluation results
  - State policies vary in how many years a teacher can perform poorly before being dismissed

- **Tenure**
  - In many RttT states, the probationary period prior to tenure has been extended
  - In some cases, tenured teachers are returned to probationary status or dismissed for poor evaluations
Teacher Turnover and Teacher Effectiveness

- Policies related to dismissals for low evaluations are based on the premise that student achievement can be improved if the least effective teachers are removed.
- Schools and districts already experience some level of teacher turnover.
  - What is the relationship between teacher effectiveness and teacher turnover?
  - Are the less effective teachers the same teachers that leave the profession?
Teacher Turnover

- The nation loses 6-13% of the teaching force annually
  - In 2007-2008, 7.6% of public school teachers moved within teaching and 8.0% left teaching altogether (U.S. Department of Education, 2010)

- Beginning teacher attrition is concerning
  - 14% leave by the end of their first year
  - 30% leave by the third year (Kirby et al., 1999; U.S. Department of Education, 2007)
  - 50% leave within five years (AEE, 2004; Pigge & Marso, 1997; Theobald & Michael, 2001)

- Intent to leave the profession remains substantial
  - In 2011, 29% of teachers indicated an intent to leave within 5 years (MetLife, 2012)

- Some schools have a harder time retaining teachers than others (Guin, 2004; Ingersoll, 2001, 2003)
Teacher Turnover and Teacher Quality

- Mixed findings regarding teacher turnover and various indicators of quality
- Researchers have found:
  - No relationship between teacher turnover and teacher quality (as measured by teachers’ ACT scores, Pre-Professional Skills Test scores, and college GPA) (Latham & Vogt, 1997)
  - The less or least effective teachers remain in teaching as compared to their more or most effective peers as measured by:
    - National Teachers Exam scores (Murnane & Olsen, 1990)
    - SAT math scores (Stinebrickner, 2002)
    - ACT scores and college selectivity (Podgursky et al., 2004)
Some research indicates no link between teacher turnover and student achievement gains (Hanushek et al., 1998). However, the predominant amount of research on teacher turnover and teacher effectiveness (as measured by student achievement) indicates the less or least effective teachers leave:

- Teachers with lower student achievement gains are more likely to leave (Goldhaber et al., 2007; Hanushek et al., 2005; Hanushek & Rivkin, 2010).
- Teachers with lower student achievement are more likely to transfer to other schools (Boyd et al., 2011).
Similar patterns are seen in beginning teacher turnover
- Teachers in their first two years who are less effective are more likely to leave (Boyd et al., 2008)

Overall, these findings suggest that the existing pattern of teacher turnover functions similarly to the intended purpose for firing policies in many of the RttT states

However, not enough evidence exists to make a claim that such patterns result in higher student achievement outcomes in schools.
Teacher Turnover: Possible Costs to Students

- **Teacher turnover patterns across schools**
  - Less qualified beginning teachers are more likely to remain in low-achieving schools (Boyd et al., 2005)
  - Teachers who stay in the same school are more effective than those who moved or left teaching altogether
    - Pattern is more pronounced in schools with more low-achieving and Black students (Hanushek & Kain, 2010)
    - This finding suggests that teacher turnover patterns as related to effectiveness may offer greater benefits to the some of the schools that struggle the most—those serving more low-achieving and Black students
Recent research adds important nuances to understanding the effect of teacher turnover on student outcomes.

- Teachers with high value-added scores leave schools that are declining in quality leading to further declines in test scores in subsequent cohorts (Chetty et al., 2011).

- All teacher turnover harms student achievement.
  - The harmful effects of teacher turnover on student achievement are larger in schools serving a greater number of low-achieving and black students (Ronfeldt et al., 2011).
Possible Costs to Students, Cont.

- **What explains the relationship between teacher turnover and student achievement?**
  - Do more or less effective teachers replace teachers who leave the profession or move to another school?
    - If a teacher in the top 20% of teachers leaves, it can take a school of average performance 6 hires to find a teacher of comparable effectiveness. In low performing schools, it may take 11 hires to find a teacher of comparable effectiveness (TNTP, 2012)
    - Teachers who leave schools are replaced by teachers who are, on average, less effective (Ronfeldt et al., 2011)
  - Teacher turnover negatively affects the students of stayers; teachers outside of the distribution (Ronfeldt et al., 2011)
Possible Costs to Teachers

- What explains the relationship between teacher turnover and student achievement?
  - Teacher turnover negatively affects teacher morale and job satisfaction
  - Teacher job satisfaction is low; the lowest in more than two decades
    - In 2011, 44% of teachers reported that they were very satisfied with teaching, a significant drop from 59% in 2009 (MetLife, 2012)
    - Job satisfaction is related to teachers’ collective efficacy (Tschannen-Moran et al., 1998)
    - Teachers’ collective efficacy is significantly related to student achievement and academic climate (Klassen et al., 2008)
    - Low job satisfaction is related to a greater likelihood of leaving the profession (Evans, 2001; Ingersoll, 2001)
Possible Costs to Schools

- **Teacher turnover, selective or voluntary, is costly**
  - Districts spend up to $17,872 replacing a single teacher (NCTAF, 2007)
  - Teacher attrition costs Chicago public schools an estimated $86 million dollars per year (NCTAF, 2007)
  - Public schools, based on average U.S. teachers’ salary, spend almost 2.6 billion dollars annually replacing teachers (AEE, 2004)

- **Schools and districts may also experience non-financial costs**
  - Loss of school community, trust among faculty
  - Higher voluntary attrition rates
Final Thoughts

- Decisions based on teacher evaluations pose many potential costs to students, teachers, and schools.
- However, many questions remain unanswered or need further research:
  - Where are financial resources, time, and energy best spent?
    - Eliminating less or least effective teachers?
    - Supporting and retaining highly effective teachers?
    - Retaining and recruiting such teachers to high-needs schools?
    - Supporting and retaining all teachers?
CONTACT INFORMATION:

ALYSON LEAH LAVIGNE, PH.D.
ROOSEVELT UNIVERSITY

ALLAVIGNE@ROOSEVELT.EDU
As schools face public opinion plus state and federal policy changes, principals and teachers are evaluated partly on student performance. The advantages have been widely touted without serious attention to the policy consequences. One criticism: This may push schools to fire teachers based on poor student evaluations. The High-Stakes Teacher Evaluation Conference will present a thoughtful analysis with an eye toward policy issues and possible negative consequences.