Misusing Standardized Test Scores to Evaluate Teachers

Jamie Levin*

I. Introduction

In September 2012, over 26,000 Chicago Public School teachers initiated one of the largest and longest strikes in United States history.[1] At the heart of thousands of picketing teachers was resistance to a new teacher-evaluation system.[2] Pursuant to Illinois' Performance Evaluation Reform Act ("PERA"), beginning in 2010 all teachers would be evaluated based in large part by their student's test score performance.[3] Illinois furthered the impact of PERA by enacting a separate provision, Senate Bill seven, which makes it possible for teachers to be fired, denied tenure, and put up for promotion based on the results of their student's tests scores.[4] Implementation of this new evaluation system comes at a time when the Chicago Public School system projects a $1 billion deficit and plans to close fifty-four schools in the area. As teachers fight to keep their jobs, the pressure to improve student scores is higher than ever. Yet, standardized tests scores do not form an appropriate basis for evaluating teachers and linking the two for high stakes purposes will have serious negative consequences for both students and teachers. Accordingly, this Note will argue that standardize test scores should not be a factor in evaluating teacher performance for high-stakes purposes in Illinois.

---

* Juris Doctor Candidate, Loyola University Chicago School of Law, Class of 2014.

[1] Strike Puts Spotlight on Teacher Evaluation, Pay, Wall Street J. (Sept. 10, 2012), available at http://online.wsj.com/article/SB10000872396390443921504577643652663814724.html.

[2] Jayette Bolinski, IL: New Chicago Teacher Evaluation at Root of Strike, Illinois Watchdog.org (Sept. 12, 2012), *available at* http://watchdog.org/56189/il-new-chicago-teacher-evaluation-system-at-root-of-strike/.

[3] Joel Hood, Public to have say on grading teachers, principals, Chi. Trib. (Nov. 19, 2011), *available at* http://articles.chicagotribune.com/2011-11-19/news/ct-met-school-performance-evaluations1119-20111119_1_part-of-teacher-evaluations-teacher-performance-chicago-teachers-union.

[4] Kerry Lester, Testing Teachers, Illinois Issues (Jan. 2013), *available at* http://illinoisissues.uis.edu/archives/2013/01/teachers.html.

The Obama administration's Race to the Top ("RTT") program is responsible for the latest push towards greater teacher accountability.[5] The RTT program holds $4 billion in funding up for grabs to states who transform their teach evaluation structures to model the "blue-print" of the RTT system.[6] The RTT program requires the use a value-added model ("VAM"), which determines an individual teacher's effectiveness by the growth in their student's standardized test scores from one year to the next.[7] Yet, there is currently no empirical evidence to prove that the VAM is an effective tool that can accurately measure a teacher's effectiveness. In fact, research reveals the model is highly unstable, unreliable and creates serious negative consequences for the classroom dynamic.

Therefore, this Note argues standardized test scores should not be used as a basis for evaluating teacher performance for high-stakes decisions. This Note will begin by addressing federal reform movements that have pressed states to adopt new evaluation measures with greater weight on standardized test scores. Next, this Note will address how the value-added model used to measure teacher effectiveness is fundamentally flawed and rife with error. Finally, this Note will discuss how the misuse of tests scores will have a serious negative impact on the classroom dynamic for teachers, as well as students.

I. Factors Pushing States to Implement Student Performance Measures

The idea of using student's test scores to measure teachers sounds somewhat reasonable since student learning is the "bottom line" of schools. Many non-educators, especially policymakers, see test scores as an easy and definitive way to measure how a teacher is performing. With this idea in mind, federal reform movements have created a great push

[5] Race to the Top: What Have We Learned from the States So Far?, AmericanProgress.org (March 2012), *available at http*://www.americanprogress.org/issues/2012/03/pdf/rtt_states.pdf.
[6] *Id.*
[7] *Id.*

towards greater teacher accountability measured by standardized tests. Yet, while this may seem like a reasonable approach to the outside eye of policymakers, to assume that standardized test scores in reading and math are indicative of what children should learn in school represents an overly simplistic vision.[8] Standardized test scores do not correlate to an effective measure to evaluate teachers. Teachers contribute to student engagement skills, personality skills, and enhanced cognitive skills, which are not shown in a standardized test. Yet, in a time of fiscal stringency, states are conforming to Obama's blueprint to get a share of the federal windfall.

a. No Child Left Behind

The landscape of education in the United States has progressively moved towards a greater focus on teacher accountability. Beginning with the passage of No Child Left Behind ("NCLB") in 2001, President Bush's policy opened a new era of testing and accountability in American public schools.[9] Schools receiving a large share of federal funds were required to be staffed by "highly qualified" teachers by 2006 and to bring all students to "proficiency" in reading and math by 2014.[10] As the stakes grew for teachers to increase student test scores, the rise and fall of scores in reading and mathematics became the central focus for teachers, principals, and school systems.[11] As a result, curriculum and funding became increasingly more centered on improving math and reading test-taking skills, while other cognitive skills took a back seat.

[8] Karen Zumwalk, Teacher Quality Can't Be Measured by Scores Alone, Columbia University Teachers College.
[9] DIANE RAVITCH, THE DEATH AND LIFE OF THE GREAT AMERICAN SCHOOL SYSTEM: HOW TESTING AND CHOICE ARE UNDERMINING EDUCATION (2010).
[10] *Id.*
[11] *Id.*

A review of over a decade of evidence demonstrates that NCLB has failed badly both in terms of its own goals and more broadly.[12] Diane Ravitch in her novel, the *Death and Life of the Great American School System: How Testing and Choice Are Undermining Education*, attributes the failure of the NCLB program to the inappropriate use of standardized testing.[13] Ravitch explains that an increase in standardized testing was not the problem, but the misuse of testing for high-stakes purposes and the belief that tests could identify with certainty which teachers and principals should be fired or rewarded.[14] Ravitch points out that tying funding to scores has encouraged states to dumb down their exams, creating illusory progress to ensure that federal dollars keep coming in.[15] Though state scores have steadily climbed since the passage of NCLB, according to National Assessment of Educational Progress the United States performance has actually steadily declined.[16] Thus, the results of the NCLB program highlight that measuring a student or teacher by a test score does not result in a positive impact on the quality of American education.

b. Race to the Top

Nevertheless, Obama's Race to the Top program, not only continues to misuse standardized tests, but now exacerbates the problem by tying test scores directly to teacher evaluations. In 2009, President Obama signed into law the Race to the Top program, making current standardized tests of basic skills more important than ever.[17] In order to win a share of the $4.3 billion dollars of American Reinvestment and Recovery Act ("ARRA") funds, states are

[12] Valerie Strauss, A decade of No Child Left Behind: Lessons from a policy failure, Washington Post (Jan. 7, 2012), available at http://www.washingtonpost.com/blogs/answer-sheet/post/a-decade-of-no-child-left-behind-lessons-from-a-policy-failure/2012/01/05/gIQAeb19gP_blog.html.
[13] *Id.*
[14] *Id.*
[15] *Id.*
[16] *Id.*
[17] *State of the States 2012: Teacher Effectiveness Policies*, N. COUNCIL OF TEACHER QUALITY (2012), *available at* http://www.nctq.org/p/publications/docs/Updated_NCTQ_State%20of%20the%20States%202012_Teacher%20Effectiveness%20Policies.pdf.

asked to develop teacher evaluation systems that take into account student test scores. Thus, the RTT program continues to misuse standardized test scores to assess teachers and further intensifies the problem by linking test scores to high-stakes purposes.

c. Impact On State Legislation

In 2009, only four states were using student achievement as an important criterion in how teacher performance was assessed.[18] As a result of the RTT program, by 2012 twenty-three states and the District of Columbia required student achievement to be a significant or the most significant factor in judging teacher performance.[19] States have taken various approaches in an attempt to gain federal grants.[20] States like Colorado, Florida, and Idaho require that student performance data, at a minimum, constitute 50% of teacher evaluations.[21] Unlike these states, the District of Columbia, Ohio, and Louisiana do not stipulate a minimum.[22]

d. Illinois Approach

Then there are states, including Illinois, which require student performance data to be a "significant factor" in evaluations.[23] Illinois defines "significant" use of student growth as at least 25% of a principal's or teacher's evaluation in the first two years of implementation, and 30% after that, with the possibility of making student growth count as much as 50%.[24] Teachers

[18] *Id.*
[19] *Id.*
[20] Bruce D. Baker et al., *The Legal Consequences of Mandating High Stakes Decisions Based on Low Quality Information: Teacher Evaluation in the Race-to-the-Top Era*, 21 EDUC. POL'Y ANALYSIS ARCHIVES 5 at 3.C (Jan. 28, 2013), *available at* http://epaa.asu.edu/ojs/article/view/1298/1043/
[21] *Id.*
[22] *Id.*
[23] *Id. See also* 105 Illinois Complied Statute Annotated 5/34-85c(a) (2011).
[24] *Id.*

not in a tested subject will be marked on the school's average reading scores, with the idea that literacy should be part of every class, from music to gym.[25]

In addition, Illinois passed a separate complimentary piece of legislation to PERA in the spring of 2011, allowing teacher evaluations to be used in decisions about tenure and layoffs.[26] Senate Bill seven requires districts to prioritize teacher's performance, instead of seniority, in layoff decisions.[27] It also makes tenure tougher, requiring high ratings on the last two years of a teacher's evaluation before the privilege is granted.[28] In addition, Illinois underperformers could be fired after two subpar evaluations, putting a greater stake in standardized test scores than a majority of other states.[29]

At the heart of the Chicago Public School strike was the use of standardized test scores for decisions regarding tenure, lay-offs, and promotions. The Chicago Teacher Union did agree with the Chicago Public School administration that the current checklist system put in place since 1967 was useless.[30] However, the Chicago Teacher Union insist that using test scores as the alternative is not the answer. To that end, sixteen different schools in the Chicago area issued a letter to Mayor Rahm Emanuel warning against standardized test-based teacher evaluations.[31] The letter outlined the Union's outrage to the new legislation, making clear that those behind the system knew little about what it is like to be in a classroom and that in order to have an effective system that improves education, policymakers must abandon their reliance on

---

[25] Sara Neufeld, *Will New Teacher Evaluations Help or Hurt Chicago's Schools?*, THE ATLANTIC (Apr. 30, 2013), *available at* http://www.theatlantic.com/national/archive/2013/04/will-new-teacher-evaluations-help-or-hurt-chicagos-schools/275415/.
[26] Lester, *surpa* note 3.
[27] *Id.*
[28] *Id.*
[29] *Id.*
[30] Neufeld, *surpa* note 25.
[31] Valerie Strauss, *Researchers blast Chicago teacher evaluation reform*, WASHINGTON POST (Mar. 28, 2012), *available at* http://www.washingtonpost.com/blogs/answer-sheet/post/researchers-blast-chicago-teacher-evaluation-reform/2012/03/28/gIQApdOfgS_blog.html

test-and-punish strategies. In addition, the letter stressed that the Chicago system was being evaluated by administrators who were not even required to receive training to guarantee consistent and accurate reviews of teachers.[32] Nevertheless, Illinois adopted the "blueprint" of the federal system that reflects policymakers ideas of what constitutes a good system, instead of what Illinois educators believe will improve the education system.

II. Criticisms of Value-Added Models

In response to the failure of the NCLB program to effectively measure a teacher based on standardized scores, the Obama administration implemented the RTT program with a requirement that teacher be evaluated using a value-added model. In contrast to the traditional methods of measuring school effectiveness under NCLB, value-added models do not look only at current levels of student achievement. Instead, such models measure each student's test score at the beginning of the school year and compares that to their score at the end of the year, and the amount the score has grown is what the teacher added.[33] The idea behind value-added modeling is to level the playing field by using statistical procedures that allow direct comparison between schools and teachers.[34] The Obama administration and other proponents of the value-added model believe that it will solve the problem delineated under the NCLB model by measuring the learning gains attributed to each individual teacher.

However, the value-added model does not solve this problem. There are confounding factors that make the link between teacher effectiveness and student achievement problematic. In addition, an important prerequisite for implementing a value-added model should be accurate

---

[32] *Id.*
[33] Harold C. Doran & Steve Fleischman, Research Matters/Challenges of Value-Added Assessment, 63 Educational Leadership 3 85-87 (Nov. 2005), *available at* http://www.ascd.org/publications/educational_leadership/nov05/vol63/num03/Challenges_of_Value-Added_Assessment.aspx.
[34] *Id.*

data linking students to teachers. However, as it currently stands, no empirical research validates the claim that value-added models are capable of differentiating between effectively and ineffectively taught students.[35] W. James Popham, a professor emeritus at UCLA and test design expert, argues without evidence, using test scores to evaluate teachers "runs counter to the most important commandment of education testing – the need for sufficient validity of evidence".[36] What research has shown is that before a teacher's performance can fairly be drawn from the VAM, three years of data are required.[37] Without this data, teachers cannot fairly and accurately be measured. Nevertheless, school systems have implemented this model without the appropriate data.

Furthermore, research has shown that the VAM model is highly unstable, ineffective, and rife with error. Research conducted by Koedel and Betts, found that only 35% of teachers ranked in the top fifth one year were again ranked in the top fifth the subsequent year.[38] This suggests that 65% of teachers actually got worse relative to their peers over a short period of time.[39] It is intuitively implausible that actual teacher effectiveness is that erratic over a short period of time. The instability of the test score means that using the VAM for any percentage of a teacher's evaluation is highly volatile. In addition, a report published by the New York Times, estimated that only three percent – or fewer – of teachers under the new systems in Florida,

---

[35] Strauss, *supra* note 31.
[36] Jack Jennings, Mind the Gap!, HUFFINGTON POST (Dec. 19, 2012), *available at* http://www.huffingtonpost.com/jack-jennings/mind-the-gap_2_b_2324262.html
[37] DANIEL F. MCCAFFREY ET AL., EVALUATING VALUE-ADDED MODELS FOR TEACHER ACCOUNTABILITY, RAND Education (2003), available at http://www.rand.org/content/dam/rand/pubs/monographs/2004/RAND_MG158.pdf.
[38] Stephan Lipscomb et al., Teacher and Principal Value-Added Research Findings and Implementation Practices, Mathematica (Sept. 14, 2010), *available at* http://www.mathematicampr.com/publications/PDFs/education/teacherprin_valueadded.pdf
[39] *Id.*

Tennessee and Michigan were read unsatisfactory, raising questions abut how effective the model will be.[40]

Another problem with a value-based system, is that a test-score gain, even using a value-added measurement, reflects much more than an individual teacher's effort. Teachers are randomly assigned to students that come with certain disadvantages and advantages, which can greatly effect their test scores. In particular, teachers with a large number of new English learners and others with special needs have been found to show lower gains than the same teachers who are teaching other students. Additionally, teachers who have a high proportion of poor students may have a harder time lifting their student's scores than teachers who work in affluent districts. This raises a large concern for Chicago Public School teachers, where roughly eighty percent of Chicago student's qualify for free or reduced lunches.[41]

In addition, the value-added model is flawed as it is impossible to fully disentangle an individual teacher from other educators that impact a student's progress. Attributing a student's test score to one specific teacher fails to take into account the work done by pullout teachers, previous teachers, specialists, tutors, and well-educated parents, all of who contribute to student outcomes.[42] Student test gains may also be influenced by school attendance and a variety of out-of-school learning experiences at home, with peers, at museums, libraries, in summer programs, on-line, and in the community. In addition, student test scores are only applicable for roughly twenty percent of teachers, including English and math teachers for grades four through eight.[43] Since only these two subjects are tested, this raises equity concerns about how all those other

[40] Neufeld, *supra* note 25.
[41] Liz Goodwin, The ABCS of the Chicago Teachers' Strike: New Evaluation system looms larges, Yahoo News (Sept. 10, 2012), *available at* http://news.yahoo.com/blogs/lookout/b-cs-chicago-teacher-strike-evaluation-system-looms-201903517.html.
[42] Kim Marshall, Rethinking Teacher Supervision and Evaluation: How to Work Smart, Build Collaboration and Close the Achievement Gap, 2nd Ed. (2013).
[43] *Id.*

teachers will be evaluated including kindergarten, first grade, second grade, art, computer, music, library, physical education, and most high school teachers.

The uncertainty and low reliability of value-added measures has caused organizations that look at the peer-reviewed research, such as RAND, to voice their concerns and caution strongly against the use of test scores for teacher evaluations. For example, RAND stated that "the research base is currently insufficient for us to recommend the use of VAM for high-stakes decisions".[44] Another report by the National Research Council's Board of Testing and Assessment concluded that VAM estimates of teacher effectiveness that are based on data for a single class of students should not be used to make operations decisions because such estimates are far too unstable to be considered fair or reliable. Despite these warnings, states continue to push through legislation that ties standardized tests to high-stake decisions for teachers.

III. Unintended Negative Consequences

As the RTT pushes states to adopt teacher evaluation systems that tie standardized test scores to teachers, the stakes have become higher then ever as teacher's livelihoods are directly impacted by the low-scores of their students. With a greater focus on standardized tests, many unintended consequences will develop that negatively impact school curriculum, student's cognitive skills, and the overall dynamic and function of the classroom.

a. Harm to Students

Linking high-stakes employment decisions to children's performance on a single test could create a truly dangerous dynamic in the classroom. The unintended but depressing effects of the testing craze was already seen under the NCLB program as art, music, and even recess

[44] DANIEL F. MCCAFFREY ET AL., *supra* note 37.

were abandoned in an effort to make more time for test prep.[45] With a focus on end-of-year testing, there inevitably will be a narrowing of the curriculum as teachers focus more on reading and math skills. This is evident, as school districts across the nation have been reducing the time available for the arts, physical education, history, civics, and other nontested subjects.

Another kind of narrowing will take place within math and reading instructional programs themselves as teachers are pressured to teach the test. High-stake tests may encourage teachers to game the system focusing on test preparation. Annual standardized exams typically include no or very few extended-writing or problem-solving items and as a result teachers that teach to the test will fail to provide a student with conceptual understanding, communication, scientific investigation, technology, and real-world applications, or a host of other critically important skills. Many critics of standardized tests have observed that when teachers focus too much on test preparation and spend their time drilling simpler, easy-to-test skills, students don't get the full college-and-career aligned curriculum to which they're entitled.[46] This also has an impact on the dynamic between students and teachers as instead of a teacher and student verses the exam, it will be a teacher versus a student's performance on the exam.[47]

In addition, as incentives for high-test scores increase, unscrupulous teachers may be more likely to engage in a range of illicit activities, including changing student responses on answer sheets, providing correct answers to students, or obtaining copies of an exam illegally prior to the test date and teaching student-using knowledge of the precise exam question.[48] A study using data from Chicago Public Schools found unusual patterns in test scores that

[45] Laura S. Hamilton & Gabriella C. Conzalez, Are High-Stakes Tests Counterproductive?, Rand Corp. (Apr. 22, 2013), *available at* http://www.rand.org/blog/2013/04/are-high-stakes-tests-counterproductive.html.
[46] Marshall, *supra* note 43.
[47] Strauss, *supra* note 31.
[48] Brian A. Jacob & Steven D. Levitt, Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating, *available at* http://sitemaker.umich.edu/bajacob/files/cheating.pdf

researchers decided could only result from cheating. Based on this evidence, researchers estimated that teachers or administrators cheated on standardized tests in at least five percent of elementary school classrooms. This situation transpired in Atlanta public schools, where more than 150 public school teachers and administrators were caught cheating on high-stakes tests. This invalidates the assessment and produces deceptively promising information on students that keep them from getting the support they need. While the vast majority of teachers may not cheat, even a small amount of it is highly problematic. Experts say the more consequences attached to test scores, the more incentives schools and teachers have to game the system.[49] Increasing the pressure by upping the stakes will only increase the negative backlash on students.

b. Harm to Teachers

As test scores become a significant factor in determining a teacher's livelihood, teachers may feel demoralized from a system that incentives a one-size-fits-all approach to teaching. Educators may be required to show how their lesson plans connect to a standard that is going to be tested and may be forced to devote more instructional time to reading and math. Teachers interested in working with students with health issues, disabilities, new English language learners, or students suffering from emotional issues may also be demoralized from working with these students because they may be penalized as student growth may not improve as much for these students.[50]

This also has a great effect on teachers working in low-income areas. Because of the inability of value-added methods to fully account for the differences in student characteristics and in school supports, teachers who work with students with the greatest educational needs will appear to be less effective than they really are. This could lead to the inappropriate dismissal of

[49] http://americanradioworks.publicradio.org/features/testing_teachers/judge.html
[50] Strauss, *supra* note 31.

teachers of low-income and minority students. This issue took center stage in the 2012 Chicago Public School strikes as some teachers felt they were being blamed for teaching in difficult, high-poverty schools. As teachers feel they are being penalized for working with low-income students, many may look for work outside low-income neighborhoods or may even leave the profession.

Furthermore, measuring student growth under the value-added model will have serious implications for teacher collaboration, which is one of the most principal tenets of good instruction since teachers need to learn from one another to solve difficult pedagogical problems. A school will be more effective if its teachers are more knowledgeable about all students and can coordinate efforts to meet students' needs. The idea of compensating teachers individually in order to differentiate their performance from their school colleagues defeats this principal. When teachers are evaluated against each other, collaboration among teachers will be replaced by competition. This effects how schools function as many teachers work as a team and share effective practices with their colleagues. Thus, increasing the stakes for standardized test scores will have serious negative consequences for students, teachers, and school systems.

IV. Conclusion

When the Chicago strike came to an end, teachers left with a deal providing them with an average seventeen percent raise over four years while lengthening the school day by about an hour. This deal may have ended the strike, but the negative consequences of the new teacher evaluation systems are still to come. As the NCLB program depicted, using standardized testing for high-stakes purposes does not improve the American education system, but rather exacerbates the problem.

In a state like Illinois with massive disparities among student levels, the new evaluation system will only cause more pressure among teachers in low-income neighborhoods to teach to the test, cheat, or leave the profession. While researchers have cautioned against the use of standardized test scores to evaluate teachers, politicians incorrectly view test scores as an appealing measurement to test a teacher's quality. In a time of fiscal stringency, Illinois has overhauled its teacher evaluation system to gain a piece of the federal funds. However, a system based on an inappropriate measure of teacher effectiveness will inevitably falter. Before it does, the likely result will be demoralized teachers, lower student achievement, and a widening achievement gap. The effect will be devastating for Illinois' education system and therefore legislatures must abandon the testing-and-pushing teacher evaluation approach.