Journal Plus Education, ISSN: 1842-077X, E-ISSN (online) 2068-1151 Vol XXI (2018), Special Issue. pp. 444-451

EVALUATING TEACHER PERFORMANCE IN THE UNITED STATES

Mindy L. Kornhaber, Ed.D.
Pennsylvania State University, USA
mlk20@psu.edu

Abstract: *Teaching is a complex task. It requires academic content knowledge, pedagogical knowledge and skills, organizational skills, understanding of human development, and interpersonal skills to engage with students, colleagues, and, in the United States, increasingly diverse families. Given its complexity, its evaluation likely needs to be multifaceted. This article first briefly presents the policy context for teacher evaluation in the United States. It then examines policies under which students' scores from standardized tests have been the essential source of data to evaluate teacher performance and describes how these evaluation systems influence teachers' classroom practice. Finally, it briefly considers evaluation systems that may better reflect, inform, and support the complex task of teaching.*

Keywords: *education system;teachers' evaluation polities;standards-based reform; teachers'practice; teachers' performance*;

1. Introduction

School systems reflect and promote their surrounding political, economic, and social systems. Given the intertwined nature of these systems, I will first describe features of these systems in the United States. Following this, I will examine how standards-based reforms (SBR) and their testing systems have been used to address and improve the fragmented, complex nature of the U.S. education system, including the evaluation of teacher performance.

I will then introduce policies of standards-based reforms, particularly the federal No Child Left Behind Act (NCLB), which governed much of public schooling between 2002-2015. Such standards-based reforms have sought to improve teacher and student performance and narrow gaps in achievement between students of different backgrounds. I will discuss the influence of these testing and evaluation systems on teachers' practice. Finally, I will consider efforts to institute evaluation systems that may better reflect and inform the complex task of teaching.

2. The Context for Teacher Evaluation Policies in the United States

The U.S. education system is complex and unequal. The complexity of the education system reflects the complexity of its political system. The Constitution of the United States, written in 1787, outlines the structure of the federal (national) government of the United States and its areas of authority or power. Powers not granted by the Constitution to the federal government was to be accorded to the states or to the people themselves. In response to the powers wielded by British monarchy which had formerly ruled the American colonies, the Constitution's framers sought to divide power and prevent a centralized ruler from exercising overarching control. At the level of the national or federal government, power was divided among the executive (Presidential), legislative (Congress), and judicial branches. In addition, power was divided between the federal government and the states. The federal government has powers over such things as minting money and declarations of war, but the Constitution does not accord the federal government power over education. Therefore, under the Constitution, authority over education was to be exercised by state governments (which

are also divided into an executive, legislative, and judicial branches) or by the people themselves.

Given the federal governments' lack of Constitutional power over education, perhaps it is not surprising that the federal government only established a Department in Education (DoE) in 1980. Among the DoE's key roles are to collect data, and enforce federal laws, including civil rights laws, that apply to any institution that receives federal money. In addition, it distributes federal money for education to states and sometimes to districts. It is the latter function that gives power to the federal role in education. In essence, the federal government exercises the "power of the purse" within the U.S. education system. However, on average, in 2015 just 8 percent of state budgets for education come from the federal government. Nearly all the remainder comes from state and local taxes (Leachman, Masterson, & Figueroa, 2017). Though federal funds are proportionately small, they have a potent effect: States' and school districts' budgets are typically stretched thin by teacher and administrator payrolls, health insurance costs, and pensions, building and maintaining school buildings, and school security, in addition to curriculum and assessment materials, educational technology, and teachers' professional development.

States and their departments of education are typically responsible for establishing curriculum standards – a responsibility for which the federal government is specifically forbidden (see, e.g., U.S. Department of Education, Laws & Guidance). They often stipulate textbooks which may be used in the state's public schools. In the era of SBR, they have also specified assessments the school districts must adopt. In addition, they establish their own requirements for teacher certification and ongoing professional development. They collect and allocate a large proportion of school tax dollars for local schools.

The day-to-day operation of the schools takes place within some 14,000 local school districts. The number of school districts varies widely across the states. Hawaii has one. New Jersey has 678. School districts are typically governed by a locally elected school board which appoints a superintendent to manage the daily operations of the district's schools. School districts must follow all state policies and federal policies regarding teacher certification requirements, data collection, and requirements regarding academic standards and teacher evaluation. However, school boards typically still retain important powers. For example, if the state doesn't determine textbooks, this is done by local school districts, or even by local teachers. School boards also have the power to hire teachers, to fire them (union rules in various states influence this as well), to establish school budgets, and requirements for professional development and teacher evaluation processes.

Across the nation's 14,000 school districts, 50.7 million public school students are educated in some 100,000 school buildings (National Center for Education Statistics, 2018). Traditionally, each school's principal was responsible for carrying out teacher observations and evaluations. Their methods for doing so varied across schools and districts. In essence, the US has never had one approach or system for teacher evaluation, certification, or professional development.

Up until the No Child Left Behind Act of 2001, many states did not require that teachers be certified or that they be certified in the disciplinary content that they actually taught in classroom. For example, in some states and school districts, "teachers" might not be licensed or certified to teach. It was also possible to have a teacher certified to teach English assigned instead to teach science or other content for which she was not certified. Such issues were especially common in rural school districts and high-poverty districts.

One implication of the U.S.'s fragmented school "system" is that it will not provide equally competent teachers across the states or even within states' very unequally funded local school districts. Not surprisingly student achievement is also quite variable, both across states and even across school districts within the same state.

3. Addressing Unequal Teaching and Learning through Standards-based Reform

To address the problems of unequal teaching and learning across states and districts, and to hold schools accountable for using public resources to benefit students, policymakers and since the 1980s have advocated standards-based reforms as have some scholars (Smith & O'Day, 1991; O'Day & Smith, 1993). By 2000, nearly every state had fully embraced such reforms. In 2001, the federal government through its powers of the purse, stipulated that all 50 states adopt NCLB, a standards-based reform policy.

Standards-based reforms are built on a theory of "alignment" (O'Day & Smith, 1993 Smith & O'Day, 1991). Specifically, each component of an education system – its instructional practices, curriculum, assessment, school resources, teacher education, and professional development – should be aligned to explicit academic standards. The standards specify what students across grades (e.g., from kindergarten to Grade 12, roughly ages 5-18) "should know and be able to do" in different disciplinary content.

To assess whether teachers were actually teaching to the state standards and students were learning the standards, states' departments of education instituted tests that, in theory, were also aligned to the state standards. To ensure that teachers and students focused their efforts on the state standards, scores from the state tests typically carried "high stakes." That is, test scores were used to assign consequences to administrators, teachers, and/or students. These high-stakes consequences varied across states and districts. For example, test results might lead to school administrators losing their jobs, teachers facing disapproval from administrators and fellow teachers, schools being closed, or students being required to repeat a grade, go to summer school or being denied a high school diploma. In some states and districts, scores might generate financial rewards for teachers themselves and/or for the school as a whole and public awards. Scores from schools and districts were also commonly published in newspapers, which subjected educators to public approval or shaming (Ravitch, 2014)

However, no state had full alignment of all the components of the standards-based reform. For example, states faced multiple obstacles with regard to aligning teacher education with the standards. The higher education systems of each state, in which teacher pre-service education takes place, are divided between public and private institutions. The latter are far less responsive to state directives. Even within the public systems of higher educations, K-12 public schools and colleges and universities have fragmented communications and rarely coordinate their efforts. Thus, pre-service teachers were not typically enabled to teach to state standards.

In most states, the components of the system that were aligned to the standards were typically tests, and then only incompletely, because standards were often too sprawling to be adequately assessed (Koretz, 2017). Curriculum and instruction also tended to be aligned more to the tests than the standards, partly because the standards themselves were too broad to test them all and because educators, schools, and/or students were judged by test scores. As the test content became clearer over years of administration, curriculum and instruction increasingly mirrored the test content much more so than the disciplinary content (Koretz, 2017). Thus, instructional time increasingly focused on learning how to take high-stakes tests, which were typically a multiple choice exam.

Although standards-based reform was aimed in part at creating more equal learning opportunities, the effects of these reforms were not uniform across schools, even within a given state's school districts using the same set of academic standards. In poorer communities, an enormous amount of the school year was spent on test preparation (e.g., Kirp, 2013). However, in wealthier districts, where students benefitted from more highly educated parents and often better-prepared and experienced teachers, the curriculum remained

enriched. As a result, SBR showed little evidence of closing achievement gaps or enabling students, across disparate districts and states to achieve at more equal levels according to federal data (e.g., The Nation's Report Card, Achievement Gaps Dashboard, n.d.).

SBR spread from the states into federal policy. Using its power of the purse, the federal government's NCLB legislation required all states to adopt standards and align their education systems to the standards. Under NCLB, all states were required to adopt standards but only for English language arts and mathematics. States were required to test all students in these two subjects each year in grades 3 through 8 (approximately ages 8-14) and once in high school. The test score results were to be used to determine how well each public school and each school district was performing. In an effort to attend to achievement disparities, test results had to be disaggregated and reported separately for different students by race, poverty, language, and disability status. Under NCLB, each school had to meet specific test score gains and to do so for all subgroups of students. Failure to meet the specified test score gains triggered a series of consequences through which schools could lose students and money, teachers and staff could be reassigned, and the school eventually closed.

Under NCLB, all students were to be proficient in both mathematics and English language arts, regardless of background poverty, race, ethnicity, native language, or disability status by the spring of 2014. That goal was both unrealistic and perhaps even nonsensical: Human performance is variable, though in an equitable education systems, achievement variation by race, gender, ethnicity or other background variables should be minimized. Universal proficiency was also absurd because there was no one set of curriculum standards across the state, in part because the federal government is not permitted to interfere in curriculum and standards (See e.g., U.S. Department of Education, Laws & Guidance). NCLB continued to enable wide variations across the states in the design of the standards themselves (e.g. how rigorous and how detailed), the testing system (e.g., the rigor of the tests and what scores represented "proficient" performance), and high-stakes consequences (e.g., states could include financial incentives and consequences for students, though these were not required under NCLB).

NCLB also required that all teachers had to be "highly qualified." Specifically, teachers should be certified to teach their subject areas. Thus, a math teacher should have taken course work and passed teacher licensing exams for math. Nevertheless, different states required different course work and licensing exams. In addition, because there are shortages of teachers, especially in rural areas, it was not possible to staff schools without exceptions to the mandate for "highly qualified" teachers. Furthermore, since salaries across school districts – even neighboring districts – can vary markedly partly because communities' tax bases vary, teachers who are certified and experienced tend to find employment in districts serving more affluent students.

While NCLB was the longest-lasting SBR, it was not the last SBR. For example, beginning in 2009, the federal government incentivized another SBR, the Common Core, which was intended to promote uniform standards and tests across the states. In addition, it was supposed to make teacher preparation and evaluation more uniform. However, political backlash to this reform largely undermined the aligned testing, which is a cornerstone of all SBR approaches (Kornhaber, Barkauskas, Griffiths, Sausner, & Mahfouz, 2017). States modified the standards and adopted different tests, continuing the prior pattern of differences in the rogor of standards and assessment across states as well as differently resourced districts within them.

4. The Influence of Standards-Based Reform on Teachers' Practice

State and federal SBR policies aimed to foster teaching and learning that reduced disparities and also to hold schools and educators publicly accountable for producing such

results. However, numerous studies, including some undertaken by the federal government (e.g., The Nation's Report Card, Achievement Gap Dashboard, n.d.) indicate that such these policy aims were largely unmet. In addition, the policies produced numerous unintended consequences.

In my view, a substantial but unexamined consequence of standards-based reforms is that it can change educators' understanding of their own and their students' obligations. That is, it can erode the understanding that teachers are professionally obligated to provide the best possible instruction to students and instead encourages educators to view students as obligated to produce good scores by which the schools and their teachers are evaluated. This problem is evident in several different influences of SBR on teachers' practice.

First, the range of curriculum offered to many students is narrowed. Subject areas that are not tested are eliminated or reduced. Such narrowing logically follows from the view that efforts to teach such subjects do not directly improve the scores in subjects that are used to evaluate teachers. Therefore, students in some schools may not have instruction in history, music, art, or even science. Since teachers must raise scores and also because they want to keep their jobs and their schools open, students can lose access to these bodies of knowledge (Ravitch, 2014; Nichols & Berliner, 2007). Students may also lose access to physical education and to recess periods. Thereby, time for these activities can be shifted to test preparation.

Second, and relatedly, within the subject areas that are tested, the range of topics may be reduced. This has occurred as the tests' format and content became clearer during years of its administration (Koretz, 2017). Thus, if a third grade math test did not include understanding and explaining information presented in graphs, then students would not be taught that information, even if they were interested and ready to do so. If poetry was not included in the sixth grade English examination, then students would not have units on poetry. In addition, the range of expression in the tested subjects was narrowed. Because the tests were overwhelmingly comprised of multiple-choice questions, it made more sense to emphasize sentence structure than literature. It made sense to teach how to eliminate a multiple-choice answer at least as much as it did to teach how to think about the relationship between time and distance. In some schools, the better part of the school year was spent on preparing for the test and teaching test taking skills (e.g., Kirp, 2013), rather than teaching a rich and full curriculum.

Third, teachers of untested subjects, such as music, history, science, or physical education, have been asked to provide instruction in subjects that are tested. This increases the time spent in preparing students to take tests and thus may increase scores and ensure that schools and teachers remain viable. It is not always evident that teachers are being directed to provide instruction in subject areas for which they lack training. For example, in a study of the state of Virginia's SBR policy, it appeared that the arts were still being taught, because there was still funding and jobs for arts teachers. However, interview data revealed that instead of teaching the arts subjects for which they were trained, art teachers were teaching vocabulary or math concepts to improve test scores (Mishook & Kornhaber, 2006).

Similarly, in some school districts and states, including New York, teachers of untested subject areas were evaluated on the basis of scores in tested subjects. That is, the scores from students they had not even taught were used in the evaluation of their teaching. The mentality behind this bizarre approach is that all teachers should be working to improve the tested performance of students, no matter whether or not they actually taught the students. This form of evaluation was judged to be arbitrary and a court has ruled in favor of the teacher who had sued New York State (Strauss, 2016).

Fourth, teachers' willingness to work with the range of learners has been undermined by SBR. Several studies, both qualitative and quantitative, showed that under various SBR policies, teachers' attention was disproportionately spent on those students whose scores were just below passing. (Booher-Jennings, 2005; Neal & Schanzenbach, 2010). In an effort to correct for this problem, a number of SBRs have embraced "value added measurement" (VAM), which evaluates educators on the basis of all students' score growth (v. for example, the percent of students who are "proficient"). Yet, VAM also presents a variety of problems (Haertel, 2013) and does nothing to eliminate the problem of narrowing of curriculum and time spent in test preparation.

Fifth, the preceding examples of the influences of evaluation systems on teacher practice illustrates that how an education policy reform that was intended to improve educational opportunity and evaluate educators' accountability can potentially de-professionalize and deskill teachers. Experienced teachers who had developed rich and extended curriculum units and dynamic instructional approaches were asked to abandon these and attend instead to test prep and score increases. Young teachers had fewer opportunities to see such rich curriculum and instruction enacted by senior colleagues. Moreover, while it is common in the U.S. for young teachers to leave the profession after just a few years, many experienced educators whose professional lives had been devoted to teaching – even those in more affluent districts – have grown increasingly dissatisfied with the profession and are considering leaving (Smith & Kovacs, 2011), even as the supply of new teachers is diminishing (Espinoza, Saunders, Kini, and Darling-Hammond, 2018).

Sixth, score-based evaluation systems have also had the unintended consequence of undermining some educators' ethics. Per "Campbell's Law" (Campbell, 1976, p. 49), whenever a quantitative indicator, such as a test score, carries important social consequences (e.g., the potential loss of one's job or one's school), the process that produces the score will be corrupted. Moreover, the result, such as a test score, will be hard to interpret. Standards-based reform is a text-book illustration of Campbell's Law. In addition to corrupting processes such as curriculum narrowing and student targeting, some educators resorted to outright cheating. This includes telling students the answers to test questions, changing students' answer sheets, encouraging other teachers to cheat, or even tinkering with the test reporting system (See e.g., McCray, 2018; Nichols & Berliner, 2007). As a result of these and many other practices that were spurred by test-based evaluation, scores produced by districts and states on the standards-based test often showed much bigger gains than audit tests of the same content (Koretz, 2017; McCray, 2018). Thus, many billions of dollars have been spent in the U.S. for test development, scoring, and reporting that do little to promote genuine improvements in learning, teaching, or teacher evaluation.

To conclude, teaching is a craft, a complex choreography of content knowledge, pedagogical knowledge and skills, interpersonal savvy, organizational skills, dedication, and care. Surely, bits of this craft can be measured. However, much of it cannot. When teacher evaluation rests largely on a test score or other single measurement, per Campbell's Law, it will undermine the process of teaching and learning and fail to provide the basis for meaningful teacher evaluation.

Because teaching is a craft, evaluating it must rely to a great extent on professional judgment. Such judgment must be cultivated and sustained in cultures that value the craft, rather than undermined by policies of test-based measurement. Such cultivation and practice existed in the British inspectorate system. This system relied on expert judgment of senior education administrators who, as staff of Her Majesty's Inspectorate (HMI), visited schools and classrooms and generated lengthy reports with suggestions for improvement. The HMI process was watered down largely to rubrics and checklists which might generate technically reliable, but hollower, feedback. Before the hard press of NCLB and a few years thereafter,

the state of Nebraska enabled school districts to develop their own standards and assessments of them. These locally-developed assessments were audited by a standardized test, and a state-wide writing assessment developed and scored holistically by teachers. Because there was local involvement and scoring by teachers, and tests were used for auditing rather than assigning consequences, evaluation was formative and useful in improving practice. Nebraska teachers valued their state's approach, despite its time demands (Dappen & Isernhagen, 2005). In addition, at Harvard Project Zero and elsewhere, "authentic assessment" has long been valued. Authentic assessment entails judging teaching and learning against real-world standards and practices for teaching and learning history, writing, mathematics and other disciplines. Each of the foregoing approaches briefly mentioned here, as well as others, reveal there are no short-cuts to good evaluation of teachers or students. The commitment U.S. policy circles to enact such short cuts have, in fact, been costly and ineffective in improving teachers' practice or students' learning.

References

Booher-Jennings, J. (2005). Below the Bubble: "Educational Triage" and the Texas Accountability System. *American Educational Research Journal*, 42(2), 231-268.

Campbell, D.T. (1976). Assessing the Impact of Planned Social Change. Occasional Paper Series, Paper #8, The Public Affairs Center, Dartmouth College. Available at http://www.wmich.edu/evalctr/pubs/ops/ops08.pdf

Dappen L. & Isernhagen, J.C. (2005) Nebraska STARS: Assessment for learning. *Planning and Changing, 36*, 3&4, pp. 147–156.

Espinoza, D., Saunders, R., Kini, T., and Darling-Hammond, L. (2018). Taking the long view: State efforts to solve teacher shortages by strengthening the profession. Palo Alto, CA: Learning Policy Institute. Available at https://learningpolicyinstitute.org/sites/default/files/product-files/Long_View_REPORT.pdf

Haertel, E. H. (2013). Reliability and validity of inferences about teachers based on student test scores. Available at www.ets.org/Media/Research/pdf/PICANG14.pdf

Kirp, D. L. (2013). Improbable scholars: The rebirth of a great American school system and a strategy for America's schools. Oxford: Oxford University Press.

Koretz, D. (2017). *The testing charade: Pretending to make schools better*. Chicago: University of Chicago Press.

Kornhaber, M.L., Barkauskas, N.J., Griffith, K.M., Sausner, E.B., and Mahfouz, J. (2017). The Common Core's promises and pitfalls from the perspectives of policy entrepreneurs and ground-level actors. *Journal of Educational Change*, 18(4), 385-412.

Leachman, M., Masterson, K., and Figueroa, E. (2017). A punishing decade for school funding. Center on Budget and Policy Priorities. Available at https://www.cbpp.org/research/state-budget-and-tax/a-punishing-decade-for-school-funding

McCray, V. (2018, October 9). With hymns and prayers, ex-APS educator reports for prison in cheating scandal. *The Atlanta Constitution-Journal*. Available at https://www.ajc.com/news/local-education/educators-convicted-aps-cheating-scandal-report-for-prison-today/tAZrJhxSzDVieSHvVjrr2J/

National Center for Education Statistics (2018). Fast Facts. Available at https://nces.ed.gov/fastfacts/display.asp?id=372

The Nation's Report Card, Achievement Gap Dashboard (n.d.). https://www.nationsreportcard.gov/dashboards/achievement_gaps.aspx

Neal, D, and Schanzenbach, D.W. (2010). Left Behind by Design: Proficiency Counts and Test-Based Accountability. *The Review of Economics and Statistics*, 92(2), 263-283.

Nichols, S. L., and Berliner, D. C. (2007). *Collateral damage: How high-stakes testing corrupts America's schools*. Cambridge, MA: Harvard Education Press.

O'Day, J. and Smith, M. (1993). Systemic reform and educational opportunity. In S. Fuhrman (Ed.), Designing Coherent Education Policy: Improving the System (pp. 250-312). San Francisco: Jossey-

Ravitch, D. (2014). *Reign of error*. New York: Vintage Books/Random House

Smith, J. E., and Kovacs, P. E. (2011). The impact of standards□based reform on teachers: the case of 'No Child Left Behind.' *Teachers and Teaching*, 17(2). Available at https://www.tandfonline.com/doi/abs/10.1080/13540602.2011.539802

Smith, M. S., and O'Day, J. (1991). *Putting the pieces together: Systemic school reform* (CPRE Policy Brief, RB-06-4/91). New Brunswick, NJ: Consortium for Policy Research in Education.

Strauss, V. (2016, May 10). Judge calls evaluation of N.Y. teacher 'arbitrary' and 'capricious' in case against new U.S. secretary of education. *The Washington Post*. Available at https://www.washingtonpost.com/news/answer-sheet/wp/2016/05/10/judge-calls-evaluation-of-n-y-teacher-arbitrary-and-capricious-in-case-against-new-u-s-secretary-of-education/?noredirect=on&utm_term=.1ecb67a2caea

U.S. Department of Education, Laws & Guidance, Elementary & Secondary Education, Part I -- General Provisions. Section 1905 (No Child Left Behind), https://www2.ed.gov/policy/elsec/leg/esea02/pg18.html