Teachers College Record
The Voice of Scholarship in Education

Exploring the Intended and Unintended Consequences of High-Stakes Teacher Evaluation on Schools, Teachers, and Students

by Alyson Leah Lavigne — 2014

Background/Context: *The stakes are getting higher for teachers daily as more and more states adopt hiring, firing, and tenure-granting policies based on teacher evaluations. Even more concerning is the limited discussion about whether or not high-stakes teacher evaluation can meet the intended outcome of improved student achievement, and at what cost. These high-stakes decisions are based on the rationale that firing ineffective teachers (as primarily measured by observation data and value-added scores) will improve student achievement. This premise is challenged by various variables and assumptions (e.g., reliability, validity, percentage fired, and turnover) that, if not met, could result in a number of possible unintended consequences.*

Focus of Study: *This paper examines the history of high-stakes teacher evaluation and the ways in which teacher evaluation data are being used in today's schools to make human capital decisions. The intended consequences and unintended consequences are explored in great detail.*

Research Design: *This paper is an analytic essay.*

Conclusions/Recommendations: *There is no evidence that high-stakes teacher evaluation can produce a more effective teacher workforce and improve student achievement. Even if basic requirements and assumptions are met (e.g., highly reliable and valid measures, retention of effective teachers, and highly effective hires), gains in student achievement may be short lived, insignificant, or practically meaningless. The possible unintended consequences could result in worse, rather than better, student achievement outcomes and increase the gap in opportunity to learn for students attending the most and least affluent schools.*

Current educational accountability practices reflect a system where schools, teachers, and students are being held accountable primarily on the basis of student achievement. Under Race to the Top, there has been an increased use or planned use of teacher evaluations to inform hiring, firing, and tenure-granting decisions. Thus, the "stakes" associated with holding teachers accountable for student learning are getting higher. The pressure on teachers to prove their worth is exacerbated by a history of criticism of American education (National Commission on Excellence in Education [NCEE], 1983; Rice, 1893; 1913). This criticism continues to be captured through various media outlets (Dillion, 2010; Foster, 2012; Stossel, 2006).

More recently, concern over American education has shifted focus to teacher evaluation. The media has documented the heated protests emerging across the nation, most notably in Los Angeles (Lopez, 2010) and New York City (Strauss, 2012). The Chicago teacher strike of September 2012 made national headlines (Banchero, Maher, Lee, & Yadron, 2012; Banchero, Porter, & Belkin, 2012; Davey, 2012) and further demonstrated the importance of this issue for teachers, students, parents, citizens, and policymakers. Op-ed pieces in *The New York Times* (Kenny, 2012) and the *Los Angeles Times* (Darling-Hammond & Haertel, 2012) continue to illustrate teacher evaluation as an important emerging concern in the educational community.

In academia, there has been significant attention placed on the statistical and methodological properties of value-added measures (Harris, 2011; Lockwood et al., 2007; McCaffrey, Lockwood, Kortez, Louis, & Hamilton, 2004; Sanders & Horn, 1998), one way many teachers are or will be measured; however, there has been growing concern about the ability of such measures to accurately and reliably capture a teacher's effectiveness (see Collins & Amrein-Beardsley, 2013, and Berliner, 2013). Little attention in the media and elsewhere has been given to the intended consequences of high-stakes teacher evaluation and decisions based on such data. In particular, it is unclear to what extent using teacher evaluations to make human capital decisions will foster teacher performance, increase student achievement, and improve the system as a whole. Of equal importance are the unintended consequences of high-stakes teacher evaluation. Although recent research has started to shed light on this topic (Amrein-Beardsley & Collins, 2012; Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012), few researchers have addressed this critical and developing issue. Thus, the goal of this paper is to a) provide a brief history of high-stakes teacher evaluation; b) demonstrate the ways in which states are using or are planning to use teacher evaluation data; c) assess the potential that high-stakes teacher evaluation can meet the desired outcomes; and d) explore the possible unintended consequences of high-stakes teacher evaluation on schools, teachers, students, and, potentially, teacher education.

ARRIVING AT HIGH-STAKES TEACHER EVALUATION

Contemporary accountability, particularly in the form of high-stakes teacher evaluation, has an interesting and evolving history (see Lavigne & Good, 2013). In this section, I will provide a brief history of high-stakes teacher evaluation. In particular, I will illustrate how models from the business world have made their way into American classrooms and the ways in which we assess teachers.

REINVENTING THE PAST?

Accountability—holding schools, teachers, and students accountable for learning—is by no means new. For example, in Ontario, as early as the late 1800s, high schools' budgets were determined by the number of students passing exams (Good, Biddle, & Brophy,

1975). In the United States, high school exit exams soon followed suit (Folts, 1996). With the 1900s, came a swift increase in the wide application of standardized testing and the scientific measurement of educational outcomes. This was combined with increased calls for results that demonstrated schools indeed were meeting intended goals coupled with consequences for not doing so. Patten (1911), an economist, argued that schools needed to provide results demonstrating societal contributions in order to acquire or maintain funding. This results-driven, test-based movement has grown significantly over the last 40 to 50 years in the United States (see Lavigne & Good, 2013).

More elaborate examples of accountability were established in the late 1900s. For example, in 1973, 27 states had some form of accountability legislation. Teacher evaluations were present in 12 states. However, states varied widely in implementation of such legislation and, in general, accountability laws and teacher evaluations were vague, laws consisted of mere recommendations, and districts were given little guidance (Good et al., 1975). The year 1983 was marked by *A Nation at Risk* in which it was recommended that "Salary, promotion, tenure, and retention decisions should be tied to an effective evaluation system . . . so that superior teachers can be rewarded, average ones encouraged, and poor ones either improved or terminated" (NCEE, 1983, p. 30). The seeds of modern-day high-stakes teacher evaluation were being sown.

BUSINESS MODELS IN EDUCATION

Corporate America drives and continues to drive results-driven ideologies into education (see Nichols & Berliner, 2007). Many have commented upon the inappropriateness of applying business models to educational practice (Callahan, 1962; Hall, 1983; Nichols & Berliner, 2007). Regardless, these models continue to be applied and used as a comparison, especially in the case of teacher effectiveness and evaluation.

As discussed below, current state-level models of high-stakes teacher evaluation, especially those that use evaluation scores to determine terminations and promotions, are not a far cry from some of the most touted employee evaluation models in business. In order to better understand teacher evaluation models, I explore what is known about such models in the business literature. Below, I will return to address the limitations of such models, providing evidence for precautions that need to be taken when considering these models as education improvement strategies.

Human Appraisal

One important component of any effective business is performance management (Murphy & Cleveland, 1995); the ability to assess workers is also known as performance evaluation systems or human appraisal systems. Within this area, three business performance evaluation models align most closely with modern-day teacher evaluation as illustrated in Race to the Top: rating scale method, ranking method, and the forced distribution method. In the rating scale method, employees are rated on performance across a series of factors. Employees are usually rated on a 5- or 7-point scale defined by a series of adjectives that describe their performance (e.g., above expectations, meets expectations, and below expectations). One of the limitations of the rating scale method is that it does not protect against score inflation, a documented problem (Bretz, Milkovich, & Read, 1992; Rynes, Gerhart, & Parks, 2005). In this absolute rating method, a significant number of employees could be rated high or low. This could result in limited variation, which is problematic if evaluations are used to inform human capital decisions and need to differentiate between employee effectiveness.

The business community has responded to this issue with other models to help curb score inflation (McBriarty, 1988) and encourage the ability of evaluators to distinguish between more and less effective employees. The ranking method is one response. In the ranking method, employees are ranked in order of overall performance. Often, this is applied within a particular group or job function rather than across an entire company. Ranking also reduces score inflation because employees cannot be equal.

Another response to inflation in rating scales is the forced distribution model. In forced distribution, employees are placed within a limited number of categories and this distribution is expected to be normal. This is probably one of the most touted methods in the business world, with at least 25% of American companies using some relative evaluation method, including Cisco Systems, Intel, Hewlett Packard, and Microsoft, to name a few (Boyle, 2001). It is also one of the most controversial.

The most publicized human appraisal model is that of former General Electric executive Jack Welch (Bossidy & Charan, 2002; Tichy & Sherman, 2001). General Electric's human appraisal system has been referred to as a "rank and yank" forced distribution model, differentiation (Grote, 2005), and a vitality curve (Welch, 2001). In the model, employees are ranked; the most effective employees are rewarded and the least effective are given time to improve or are eligible for termination. Until the mid-2000s, General Electric implemented a 20/70/10 split. Raters identified their top 20%, middle 70% (the "vital" middle), and bottom 10% of workers (Kwoh, 2012). Welch (2001) argued, "Year after year differentiation raises the bar higher and higher and increases the overall caliber of the organization." (p. 158). He also believed that the top performers should receive raises two to three times that of the next level of performers. Proponents of forced distribution models argue that such models motivate the best employees, force managers to complete honest evaluations, support strong leadership, cultivate a climate of meritocracy, and eliminate ineffective workers (Boyle, 2000; Jenkins, 2001; Welch, 2001).

Lessons From the Past: Forced-Distribution Models

Forced-distribution models have been swiftly adopted in business, but have been done with a paucity of empirical evidence (Pfeffer & Sutton, 2006; Scullen, Bergey, & Aiman-Smith, 2005). Further, employees at major companies (e.g., Microsoft, Conoco, Ford, and Goodyear) have filed class-action lawsuits claiming that these evaluation models result in discrimination (Scullen et al., 2005). This, in combination with extensive attention in the media (Clark, 2003; Jenkins, 2001) and elsewhere (Meisler, 2003; Scullen et al.,

2005), illustrates the controversial nature of forced-distribution models. Some have even suggested that criticism has led to a decreased use of these models (Kwoh, 2012). Emerging criticisms have also prompted a significant exploration into the limitations of such models.

In one study, Scullen et al. (2005) examined, though simulation, the potential for an organization to improve performance over time by firing workers and replacing them with more promising applicants. The 30-year simulation tested numerous models that varied on firing rate, inter-rater reliability, model validity, selection ratios (ratio of people hired to number of applicants), and voluntary turnover rates. Findings indicated that although all models improve performance over time, early loss in gains could be attributed to high voluntary turnover. Some models experienced their first decrease at 3.5 years. Thus, forced-distribution models may improve overall performance, but the benefits are short-lived. Hence, the idea that models that eliminate and replace the most ineffective employees, as a means of continually raising performance and workforce quality, is not a realistic or plausible long-term goal.

Corporate America has had its fair share of high-stakes evaluations and many of the most aggressive models have been scaled back. The failures and challenges faced in the business world's version of high-stakes evaluation paints an uncertain picture for the application of such models in education. The history of high stakes in previous educational efforts offers additional concerns. Previous educational reforms, such as the No Child Left Behind (NCLB) Act, intended to close the achievement gap by holding states and individual students accountable for student learning. As with NCLB, valid concerns were raised and later affirmed about the ability of high-stakes testing to meet intended outcomes of closing the achievement gap (Amrein & Berliner, 2002). Further, the unintended consequences of high-stakes testing are well documented, including a narrowed curriculum, cheating, teaching to the test, and lower standards for students (Jones, Jones, & Hargrove, 2003; Nichols & Berliner, 2007; Orfield & Kornhaber, 2001; Payne, 2008; Ravitch, 2010; Valenzuela, 2005). Nichols and Berliner (2007) argued that Campbell's Law explains these outcomes. Campbell (1976) argued, "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" (p. 49). Given the growing emphasis on the high-stakes nature of teacher evaluation, reasoning would follow that similar unintended consequences may occur. Taking lessons both from the business world and from previous educational reforms, this brief history of high-stakes teacher evaluation provides concerns in the ability of modern-day teacher evaluation to achieve its intended outcomes, while minimizing negative unintended consequences.

CONTEMPORARY TEACHER EVALUATION

In the second half of the paper, I address the intended and unintended consequences of high-stakes teacher evaluation. Prior to doing so, I explore the extent to which data from teacher evaluations are currently being used. For example, are evaluation results made public? What are the consequences if a teacher receives a low evaluation? Will he or she be fired? Will teachers with high evaluations receive monetary compensation? Are different policies in place for tenured and non-tenured teachers? Essentially, I explore how and to what extent states have hiring, firing, and tenure policies in place that are based on teacher performance. If Campbell's Law holds, the severity of the "stakes" is crucial.

In examining all 50 states and the District of Columbia, in 2011, 20 had policies that require teachers to be eligible for dismissal based on evaluation results. In some states, this policy expands to all layoffs, including ending "last in, first out" policies and using performance as the first data point for dismissal. In states like Florida, tenure has been effectively eliminated and annual contracts are renewed based on teacher performance (National Council on Teacher Quality [NCTQ], 2011). In some cases, policies allow for teacher evaluations to be tied to human capital decisions, but these policies may or may not be enacted (TEAM Questions Correspondent, personal communication, October 1, 2012).

RACE TO THE TOP TEACHER EVALUATION

Accountability in education has experienced a shift in focus from primarily schools to, now, teachers. This is most apparent in Race to the Top, a $4 billion dollar program that awards states with coherent and rigorous plans to prepare students for success in college and the workplace and to compete in the global economy; build data systems that measure student growth and success and inform instruction; and recruit, develop, reward, and retain effective teachers and principals; and finally, turn around the lowest-achieving schools (U.S. Department of Education, 2012).

To a great extent, Race to the Top has been leading the push for using teacher evaluations to make human capital decisions. As outlined in the Race to the Top application, 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)

Hence, I used Race to the Top winners to examine, more closely, the ways in which states are using or are proposing to use evaluation data to make decisions about teachers and their career trajectories. I focus specifically on firing policies and preliminary or initial evaluation results (see Lavigne & Good, 2013, for information on other human capital decisions). It is important to note that Race to the Top winners vary significantly in stage of implementation due, in part, to the phase in which they were granted Race to the Top funds (see Table 1).

Table 1. Race to the Top Winners

Phase 1
Delaware
Tennessee
Phase 2
District of Columbia
Florida
Georgia
Hawaii
Maryland
Massachusetts
New York
North Carolina
Ohio
Rhode Island
Phase 3
Arizona
Colorado
Illinois
Kentucky
Louisiana
New Jersey
Pennsylvania

(U.S. Department of Education, 2010, 2011)

In order to provide a more precise illustration of teacher evaluation and its uses, I chose to examine only states within Phase 1 and Phase 2. The selection pool criteria were set in order to examine teacher evaluation models that were beyond the planning stage and either were in the process of being piloted or were implemented. From these 11 states and the District of Columbia, I chose two, the District of Columbia and Tennessee (see Lavigne & Good, 2013, for additional Race to the Top profiles). These two were chosen because they differ in terms of the "stakes" for teachers. Both have made significant progress in implementation and have initial and preliminary results available. Hence, these two also were chosen because the preliminary results from those leading in implementation provide important learning opportunities for the future of high-stakes teacher evaluation.

District of Columbia

The District of Columbia Public Schools' IMPACT system, implemented in 2009, bases teacher evaluations on three components: student achievement, teaching and learning (observation), and commitment to the school community. When all data are available, 50% of a teacher's evaluation is based on student achievement data and the remaining two components account for the remaining 50% of a teacher's evaluation score (District of Columbia Public Schools [DCPS], 2012a).[1] For general education teachers without value-added data, 75% of their evaluation is composed of observation data, 15% is teacher-assessed student achievement, and 10%

is commitment to the school community (DCPS, 2012c). The fact that not all teachers will be evaluated using the same measures (or with different weights applied to the same measures) is hugely important, understudied, and could create a significant amount of error and bias because measures may vary widely in their reliability and validity. Teachers receive a final overall IMPACT rating on a five-point scale: highly effective, effective, developing, minimally effective, and ineffective. Highly effective teachers are eligible for annual bonuses (up to $5,000) with additional opportunities for base salary increases because of advancement to the Leadership Initiative For Teachers (LIFT) career stage (DCPS, 2011a). Additional performance-based pay opportunities are available to Washington Teachers' Union members through IMPACT*plus* (DCPS, 2011b). Teachers who are rated minimally effective are encouraged to take advantage of professional development opportunities. If a teacher is rated minimally effective for two consecutive years, he or she is eligible for dismissal. In addition, the teacher's salary is frozen until a rating of effective or higher is earned. Teachers who receive an ineffective rating are subject to dismissal.

The results from the first two years of implementation in the District of Columbia Public Schools are outlined in Table 2.

Table 2. District of Columbia Public Schools Teacher Evaluation Results: 2 Years of Implementation

Rating	2009-2010	2010-2011
Highly Effective	1499 (23%)	1213 (18%)
Effective	4086 (62%)	4269 (65%)
Minimally Effective	727 (11%)	750 (11%)
Ineffective	135 (2%)	113 (2%)
Ineligible to Score	143 (2%)	238 (4%)
Total	6590	6583

As demonstrated in Table 2, at the conclusion of the 2009–2010 school year, 23% of teachers were rated as highly effective, 62% as effective, 11% as minimally effective, and 2% as ineffective. Hence, the 135 or 2% of teachers who were rated as ineffective were given separation notices with the option to resign (or retire, if eligible) prior to separation. At the conclusion of the 2011 school year, 113 teachers were given separation notices in addition to 175 teachers who were rated as minimally effective 2 consecutive years. Thus, since implementation, DCPS has dismissed a total of 423 teachers for poor performance (DCPS, 2011c).[2]

Tennessee

Tennessee, a long-time leader in value-added assessment, was one of the first states to implement a statewide evaluation system. The Tennessee Educator Acceleration Model (TEAM) was implemented statewide in the 2011–2012 school year. Under their First to the Top Act and as outlined in TEAM, 50% of a teacher's evaluation is calculated using student achievement data. Of that, 35% is accounted for using student growth as represented by value-added scores or a comparable measure, and the remaining 15% is additional measures of student achievement adopted by the State Board of Education and chosen through mutual agreement by the educator and evaluator. The remaining 50% of a teacher's evaluation is composed of observation data, personal conferences, and review of prior evaluations and work. Teachers receive a score on each of the three measures (e.g., student growth, student achievement, and observations) using a 5-point scale. Using the above percentages and teachers' scores on each of the three measures, a final effectiveness score is calculated using a 5-point scale: significantly below expectations, below expectations, at expectations, above expectations, and significantly above expectations (Tennessee Department of Education [TDOE], n.d.a).

In Tennessee, although evaluation scores can be used to inform human capital decisions, it is not a requirement as designated in policy or statute. Currently, evaluation scores are not being used to inform dismissal; however, they will be used to inform tenure. Additionally, tenure legislation has been changed alongside implementation of the TEAM. Teachers are now eligible for tenure after five years if they have received a rating in the top two evaluation categories in the last two years of their probationary period. 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. Teachers who do not meet the second criteria can continue teaching under their current contract (TDOE, n.d.b).

In July 2012, the Tennessee Department of Education released data from the Tennessee Value-Added Assessment System [TVAAS] and observation ratings from Year 1. See Table 3.

Table 3. Tennessee's Distribution of TVAAS and ObservationScores from 2011-2012

Level	1	2	3	4	5
TVAAS Individual Teacher Effect	16.5%	8.1%	24.5%	11.9%	39.1%
Observation	0.2%	2.2%	21.5%	53.0%	23.2%

Table 3 includes the results from the 2011–2012 school year. Two rows are listed: the individual teacher value-added score (TVAAS) and the observation score. The percentage of teachers receiving each rating for the two different measures is listed. As illustrated in Table 3, the largest percentage of teachers (39.1%) was rated as a "5" or significantly above expectations in the value-added score. On the other end of the spectrum, 16.5% of teachers were rated as a "1" based on the same measure. It is important to note that only a portion of all teachers in Tennessee have an individual teacher value-added score. The fact that distributions will likely vary significantly for relative and absolute evaluations is incredibly important. For teachers who have value-added scores, approximately half of teachers will be above and below the mean; however, teachers who do not have value-added data may be at an advantage with the higher scores that have been recorded on observation instruments (Sawchuk, 2013). The second row represents the distribution of observation scores across the 5-point scale. As demonstrated in Table 3, the largest percentage of teachers (53%) received a rating of "4." Only .2% received a rating of "1" (TDOE, 2012a).

Those wary of value-added measures may argue for using measures in a low-stakes fashion (Darling-Hammond et al., 2012) or in combination with multiple measures to assess a teacher's effectiveness. The rationale behind the latter recommendation is that multiple measures allow for a greater opportunity to reduce bias and increase criterion validity. The differences in the distributions illustrated in Table 3 raise some concerns. Qualitatively, the Tennessee Department of Education in the Year 1 report indicated that the average observation score for a teacher with an individual value-added score of "5" was just above "4." These results suggest that evaluators are able to identify their higher performing teachers. However, teachers with a value-added score of "1" received an average observation score of 3.64. The Tennessee Department of Education concluded that this demonstrates an inability or unwillingness on the part of evaluators to identify the lowest performing teachers (TDOE, 2012a). In response, the Tennessee Department of Education will be offering additional support to help strengthen the consistency between the qualitative and quantitative data. They have also implemented consequences for districts that do not demonstrate appropriate alignment between scores and distributions. Districts that do not demonstrate results within the acceptable ranges could lose their ability to apply for TEAM flexibility, which allows districts to have more local control by implementing adjustments to the evaluation system (TDOE, 2012b). However, low correlations between observations scores and TVAAS could be due to a variety of factors related to measure, bias in the observers, or instability in teaching practices, to name a few (see Collins & Amrein-Beardsley, 2013, Berliner, 2013, Good, 2013, and Herlihy et al., 2013 for greater elaboration on alternative hypotheses). Regardless, preliminary results from Tennessee provide an important opportunity to explore why distributions in qualitative and quantitative measures vary so widely and do so more for some teachers than others.

In sum, the models above illustrate just two examples of Race to the Top teacher evaluation. Some states are moving forward swiftly in using scores to inform human capital decisions. Many states are simultaneously changing tenure policies with teacher evaluation. Value-added modeling simulates, to some extent, a forced-distribution model in that teachers are scored in relative ways to one another and around a grand mean. These scores, however, are only part of an overall evaluation score that is often composed of both relative and absolute scoring methods. Regardless, there are recommendations and consequences in place that support an underlying assumption that there *will* and *should* be a certain percentage of teachers who are rated as ineffective and, in some cases, dismissed (TDOE, 2012b). Before discussing the possible *unintended* consequences of teacher evaluation models, I first turn to explore the potential of such models to achieve the *intended* outcome of improved student achievement.

INTENDED OUTCOMES OF HIGH-STAKES TEACHER EVALUATION

Other papers in this special issue address characteristics related to the ability of high-stakes teacher evaluation to accomplish the intended outcomes (see Berliner, 2013). Hence, my discussion of this topic is brief and provides one of many perspectives on this topic.

Teacher evaluation serves multiple purposes. One is to identify and provide support for struggling teachers. *High-stakes* teacher evaluation, however, serves additional needs. It allows districts to identify the effectiveness of teachers in order to inform dismissal and layoff decisions, removal and grant of tenure, promotion, and bonuses. These are also the same "ends" described above in Race to the Top. These are also the *minimum* suggested uses of teacher evaluations. One assumption inherit in the goal of high-stakes teacher evaluation is that removing ineffective teachers will increase student achievement by improving the overall quality of the teacher workforce. From a theoretical standpoint, I address a few variables that could undermine or support this assumption. It is important to note that success, under Race to the Top, has not been concretely defined. It is unclear how much improvement in student achievement is needed in order for high-stakes teacher evaluation to be deemed successful. Does

improvement need to occur across all subjects? Grades? Groups of students? Should improvement be consistent across time? Amidst these unanswered questions and with no precise outcome, I assess the potential that high-stakes teacher evaluation can improve student achievement.

ASSESSING POTENTIAL: ONE APPROACH

In this approach, I use the simulation model tested by Scullen et al. (2005) as a framework. I highlight four variables that interact simultaneously as determinants of evaluation policy "success": percentage fired, reliability, voluntary turnover, and hiring pool. See Berliner (2013) for a discussion of validity.

Percentage fired

For a case in point, assume that evaluation models represent a similar distribution, as outlined at General Electric (10% of workers are rated ineffective). Organizations have the option to implement policies that require all 10% (or 5%, 2%, and so on) of ineffective employees be fired. Hence, the percentage of employees fired can determine improvement and, more importantly, how much improvement. However, in the business literature on forced-distribution, limited evidence exists about what percentage of the workforce needs to be dismissed annually in order to achieve substantial gains. It is hypothesized that firing a larger percentage of the workforce, with all other variables equal, would offer the largest gains. In a 30-year simulation, this hypothesis holds true. Declines in growth happen later in time when the percentage fired is higher. It is believed that a larger percentage fired allows organizations to more quickly adjust to hiring or firing errors (Scullen et al., 2005).

There is no designated percentage of teachers who should be fired in many of the Race to the Top states that plan to or are using teacher evaluations to dismiss ineffective teachers. Although there is the expectation that *some* teachers should be rated ineffective (TDOE, 2012a; 2012b), it is unclear what percentage of teachers would need to be fired to improve student achievement and the extent of the improvement. In DCPS, 2–4.5% of the teachers have been dismissed for two consecutive years (DCPS, 2011c). Assuming that schools and districts vary in their distribution of teacher effectiveness, this percentage should also vary. Whether or not the percentage dismissed should stay stable over time is a point of contention that I return to below.

Furthermore, in current Race to the Top state models, there are few specifications about the distribution of ineffective teachers eligible for dismissal (see Lavigne & Good, 2013). For example, is there a limited (or expected) number of teachers eligible for dismissal in a given school? Grade? If such ratings are relative and raters are explicitly or implicitly expected to produce a particular distribution, a teacher in a "good" grade or school could be eliminated, but that same teacher may be retained in a "poor" grade or school, especially if measures in teacher evaluation cannot accurately control for such variables (see Collins & Amrein-Beardsley, 2013, and Berliner, 2013). However, if distributions were applied school- or district-wide, teachers from an entire grade or school could be eligible for dismissal. Hence, the relative and absolute nature of teacher evaluations may be just as important as the actual percentage dismissed (Roch, Sternburgh, & Caputo, 2007).

Reliability

Firing ineffective teachers as a school improvement strategy can only be as effective as the actual evaluations used to assess teacher performance. High levels of reliability are important, and even more so if the measures are being used for high-stakes decisions. In order to do so, the following needs to be established: are teachers' evaluation scores reliable across time and across raters? Teacher evaluations are primarily determined by observation data and student achievement data. Hence, it is valuable to establish the stability of teaching practices across time and teaching assignments (see Good, 2013) and the stability of teachers' student achievement gains across teaching assignments, time, and measures (see Collins & Amrein-Beardsley, 2013, for value-added scores). Beyond these two issues, high inter-rater reliability can reduce the chances of error and bias on observation measures. A meta-analysis by Viswesvaran, Ones, and Schmidt (1996) indicated that the mean inter-rater correlation for a supervisor's ratings of employees is .52. Given that multiple ratings across multiple raters will likely reduce error or bias, the lower the inter-rater reliability, the greater likelihood of error. In a report from the Measures of Effective Teaching Project, the highest reliability (.67) can be achieved when teachers are observed four different times by four different observers (Bill & Melinda Gates Foundation, 2012). It is important to note that although an observation instrument may have high reliability, it may not necessarily be valid. Observers may be able to reach high levels of agreement on observation instrument items that are not good indicators of an effective teacher, or worse, measure other factors entirely. Further, all instruments vary in their reliability, validity, and use as part of a composite score (see Lavigne & Good, 2013). Hence, the level of actual reliability within these three domains will determine the accuracy by which such policies can identify ineffective teachers for dismissal and the subsequent success of high-stakes evaluations to meet intended outcomes. Additionally, the extent to which states put in place practices that support high reliability (e.g., many observations by multiple unbiased raters and quality training) are important (see Herlihy et al., 2013).

States have been struggling to maintain standards outlined in initial teacher evaluation plans. In many cases, teacher evaluations fall under the responsibility of the school administrator. When teacher evaluations double or even triple in frequency in a given year, incredible strain is placed on administrators' time. Tennessee has experienced the practical challenges of putting such plans into place. Policy changes were passed in Year 2 that would help reduce the number of observations and increase the strategic use of observations. For example, teachers with a "1" for an overall or TVAAS will be required to have four classroom observations the following year. Teachers scoring "5" overall or for TVAAS will be required to have one classroom observation and two walk-throughs (TDOE, 2012c). It is unclear how much reliability will be sacrificed with these changes (see Herlihy et al., 2013). These changes also assume that teachers' practices and value-added scores are stable and that value-added measures are more accurate representations of a teacher's effectiveness, which is an assumption that cannot easily be supported by research (see Good, 2013, and Collins & Amrein-Beardsley, 2013). The financial and practical limitations are real. The ability of states to fund multiple

observations by multiple raters is nearly impossible. There are serious limitations to states' abilities to meet the necessary standards of reliability. These limitations are even more alarming if evaluations are being used in a high-stakes manner.

Voluntary Turnover

Beyond factors related to the evaluation measures, the workforce plays an important role in determining how human capital decisions will support organizational improvement. For example, all industries and professions experience some level of turnover. The percentage of voluntary attrition and who chooses to leave are important factors in determining the effectiveness of high-stakes evaluation. If ineffective employees can be identified accurately and reliably, are fired, and then replaced by more promising applicants, in theory, the workforce would improve in quality over time. Assuming this occurs, then the bar is essentially raised and the quality of the workforce improves over time. Unfortunately, this also indicates that those leaving the profession are of even higher quality, and replacing them with comparable or better hires becomes increasingly difficult. Over time, these voluntary turnovers become even more damaging and detrimental to organizational improvement. Scullen et al. (2005) illustrated in a simulation study that this holds true; gain over time is attributed more greatly to an organization's (or school's) ability to retain highly effective employees than to other factors such as percentage fired (Scullen et al., 2005).

In education, research on turnover in the last decade has shifted from understanding who leaves to the *effectiveness* of teachers who leave the profession (Chetty, Friedman, & Rockoff, 2011; Goldhaber, Gross, & Player, 2007; Hanushek, Kain, O'Brien, & Rivkin, 2005; Ronfeldt, Lankford, Loeb, & Wyckoff, 2011). In particular, it would be important to determine whether or not this phenomenon happens naturally—that less effective teachers are the same teachers who leave the profession and effective teachers are retained. If this is indeed the case, this voluntary turnover could make high-stakes teacher evaluation unnecessary or less effective (Scullen et al., 2005).

Researchers who have examined teacher effectiveness, as measured by student achievement outcomes, have demonstrated that teachers with higher student achievement gains are more likely to be *retained* (Boyd, Lankford, Loeb, Ronfeldt, & Wyckoff, 2011; Goldhaber et al., 2007; Hanushek et al., 2005; Hanushek & Rivkin, 2010) and less likely to transfer to other schools (Boyd et al., 2011). A similar pattern is found in beginning teachers. Teachers in their first two years who are more effective are more likely to stay (Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2008). This finding holds true when using principal's evaluations as a measure of effectiveness (Murnane, 1984). This would suggest that the existing pattern of teacher turnover functions like the firing policies in many of the Race to the Top states; more effective teachers remain, less effective teachers leave. However, it is inconclusive whether or not these patterns resulted in higher student achievement outcomes. Some important questions remain: Who replaced the teachers who left? Were the replacements more or less effective? How much more? Further, although it appears natural patterns of retention support school improvement, this assumption can only be met if the teachers who remain maintain their effectiveness across time.

Hiring Pool

As mentioned above, the effectiveness of the replacements is just as important as those fired. If an organization is unable to attract and hire employees of comparable or greater effectiveness, improvement is unlikely. In the worst case scenario, ineffective employees are fired and the potential replacements available in the hiring pool are less effective. The measure of quality of the hiring pool in business is often measured by a variable called the selection ratio, the number of people hired divided by the number of applicants. There are standard selection ratios that are determined to be "good" in business (Scullen et al., 2005).

In education, there exists a significant amount of knowledge about the hiring pool, and schools vary significantly in their ability to both retain and hire highly effective teachers. Also, the teacher workforce has changed significantly over time. There are multiple questions that need to be addressed to determine how the hiring pool would affect the success of high-stakes evaluation policies in education. For example, if the least effective teachers are fired, are there enough replacements? Assuming there are indeed enough replacements, who would be replacing those fired? The demographics of the teacher workforce have changed substantially over time, particularly in terms of experience. In 1987–1988, the modal teacher had 15 years of teaching experience; experience represented a normal distribution. However, in 2007–2008, this number was 1—the teaching workforce was made up of nearly 200,000 first-year teachers (Ingersoll, Merrill, & Consortium for Policy Research in Education, 2012). Assuming that new teachers represent a large number of the hiring pool, it can be anticipated that those fired will be replaced by a beginning teacher. Furthermore, in studies examining teacher turnover, replacements are, on average, less effective than teachers who left (Ronfeldt et al., 2011). Both of these findings suggest that high-stakes teacher evaluation policies may result in students being taught by less effective teachers and teachers with fewer years of experience.

This approach offers one framework to assess the ways in which high-stakes teacher evaluation can improve student achievement. All of these four variables simultaneously interact. For example, an unreliable measure may impact both the firing of ineffective teachers and the hiring of effective ones. And, all measures need to be both reliable and valid (see Berliner, 2013). Furthermore, if schools are losing their best teachers at the same rate in which they are firing ineffective teachers, they may fail to make gains. Many states have policies in place to retain highly effective teachers. However, such policies may be insignificant if the measures by which teachers are evaluated fail in face validity and result in increased attrition of such teachers. Beyond these basic theoretical considerations, unintended consequences have the potential to undermine the effectiveness of high-stakes teacher evaluation; create additional costs to schools, teachers, and students; and, in the end, create a less effective teaching force rather than a more effective one.

UNINTENDED CONSEQUENCES OF HIGH-STAKES TEACHER EVALUATION

As mentioned above, limited research exists on the potential unintended consequences of high-stakes teacher evaluation. Using the existing literature, I describe below only a few potential unintended consequences of high-stakes teacher evaluation. In many cases, these outcomes overlap significantly with those described above.

TEACHER ATTRITION: A CONTINUATION

Above, it was established that the natural patterns of attrition in education function like high-stakes teacher evaluation policies in that the least effective teachers leave teaching and the most effective remain. Recent research adds important nuances that illustrate the complex nature of teacher turnover in the context of high-stakes teacher evaluation. In particular, it is important to establish if patterns of turnover function differently across schools. In Texas, teachers who stayed in the same school were more effective than those who left. These findings were more pronounced in schools with more low-achieving and Black students, indicating that the difference between the effectiveness of stayers and leavers was significantly larger in these schools than others (Hanushek & Rivkin, 2010). This latter finding suggests that teacher turnover patterns may offer *greater* benefits to schools that struggle the most—those serving more low-achieving and Black students. However, Chetty et al. (2011) found that teachers with high value-added scores leave schools that are declining in quality and this departure leads to further declines in tests scores in subsequent cohorts. Hence, the effects of teacher turnover are most harmful to schools struggling with low student achievement. However, other research indicates that *all* teacher turnover negatively affects student achievement (Ronfeldt et al., 2011). Students in grade levels with higher turnover do worse than years when turnover is lower in both English language arts (ELA) and math. Student math scores were 6–7% of a standard deviation lower in years when there was 100% turnover as compared to years when there was no turnover. It was also found that the negative effect of teacher attrition on student achievement was larger in schools serving a greater number of low-achieving and Black students. This effect was two to four times larger in ELA and approximately two times larger in math.

Taken together, these findings have important implications for selective attrition in the form of high-stakes teacher evaluation. First, as mentioned above, if there are no restrictions on how dismissals are distributed, teachers of an entire grade could be dismissed. According to Ronfeldt et al. (2011), this would have severe consequences for student learning. Second, policymakers should be concerned about how high-stakes teacher evaluation and related firing policies may be counterproductive and harm student achievement rather than help it. Why this might occur could be due to a number of factors described above (e.g., effectiveness of replacements); however, such factors do not fully explain the relationship between teacher turnover and student achievement (Ronfeldt et al., 2011). Third, such policies may exacerbate achievement differences between students in affluent and less affluent schools.

TEACHER STRESS, JOB SATISFACTION, MORALE, AND LOCUS OF CONTROL

Expanding on the research done by Ronfeldt et al. (2011), there are numerous hypotheses as to why teacher turnover is harmful to student achievement. One explanation is that high attrition rates affect teacher morale and job satisfaction, which in turn result in lower student achievement outcomes. It is also possible that high-stakes teacher evaluation may exacerbate the relationship between these variables and support teachers' endorsement of external locus of control, another factor related to lower student achievement outcomes. Loss of trust between faculty and administrators may provide additional strains in an already challenging and criticized profession.

Kyriacou (2001) described teacher stress as negative emotions resulting from a teacher's work. One-quarter of all teachers report that teaching is a very stressful job (Kyriacou, 2001). Teachers with greater teacher stress have lower self-efficacy (Betoret, 2006; Schwarzer & Hallum, 2008; Skaalvik & Skaalvik, 2007), poorer teacher–pupil rapport, and lower levels of teaching effectiveness (Abel & Sewell, 1999; Kokkinos, 2007). In the context of high-stakes teacher evaluation, teachers may experience greater stress from their evaluations being high-stakes, particularly if they feel such evaluations are unfair or unjust. Even teachers who feel confident in their teaching, particularly if they perceive no significant differences between teachers terminated and those offered continued employment or tenure, may experience an increase in stress because they may perceive themselves at equal risk of termination (Folger & Konovsky, 1989).

Stress and teacher burnout, highly related variables, are also related to locus of control. Teachers and student teachers with an internal locus of control report lower stress than those with external ones (Huston, 1989; Lunenberg & Cadavid, 1992; Sadowski & Blackwell, 1985, 1987). This is particularly true for teachers in the United States (Crothers et al., 2011). If teachers believe that teacher evaluations are not a reliable and valid measure of their teaching (see Berliner, 2013, and Collins & Amrein-Beardsley, 2013), they may attribute their evaluations to luck rather than their own capabilities. Teachers' locus of control will be external. Helplessness is one coping strategy when individuals are faced with both high levels of stress and an external locus of control (Cohen, Rothbart, & Phillips, 1976). Teachers are less likely to engage in improving teaching practices if they do not believe such practices result in better student outcomes. These beliefs and practices are related to lower levels of student achievement (Fang, 1996; Weiner, 1985). The relationship between these two variables may be further strained if both administrators and teachers are not able to connect observation data to student achievement data in meaningful ways. Tennessee's low correlations between observation scores and student achievement data are one example of this (TDOE, 2012a).

TEACHER RETENTION

Highly Effective Teachers

As mentioned above, attrition of highly effective workers can limit the ability of any evaluation model to improve organizational performance. Further, the research above on evaluation policies, job stress, and, subsequently, turnover suggests that high-stakes

teacher evaluation may increase voluntary attrition. In particular, low job satisfaction is related to a greater likelihood of leaving the profession (Ingersoll, 2001). This is particularly problematic in the case of highly effective teacher attrition. When a teacher in the top 20% leaves (teachers who helped students learn two or three additional months' worth of math and reading compared with the average teacher), 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 (The New Teacher Project [TNTP], 2012). The negative unintended effects of these consequences will disproportionately affect schools that struggle the most to attract and retain teachers, schools that serve a larger percentage of minority students and students in poverty (Cloudt & Stevens, 1995; Guin, 2004; Ingersoll, 2001; Shen, 1991). This is one explanation as to why teacher turnover is most harmful to such schools (Ronfeldt et al., 2011). These findings also raise important questions regarding policies that allow highly effective teachers to move out of the classroom to mentor teachers. What is the gain in student learning that can subsequently be attributed to the mentor teacher's mentees? What is the loss accrued if the mentor teacher is replaced by a less effective teacher?

Beginning Teachers

In the United States, 14% of new teachers leave by the end of their first year. By the fifth year, 50% have left the profession (Alliance for Excellent Education [AEE], 2004; Ingersoll & Perda, 2012; Pigge & Marso, 1997; Theobald & Michael, 2001). The attrition rates of first-year teachers have increased 34% from 1988 to 2008 to a rate of 13.1% (Ingersoll et al., 2012). Given the changes in the demographics of the teacher workforce, there are more beginning teachers now than ever. Furthermore, these same teachers are leaving at incredibly fast rates. Some have argued that during the first years of teaching, teachers develop their capacity for improving student achievement. This ability begins to stabilize in the fifth year of teaching (Darling-Hammond, 1999; Johnson, Berg, & Donaldson, 2005). These findings have been recently replicated using value-added data (Kersting, Chen, & Stigler, 2013). Given that beginning teachers are still developing and are growing in their ability to acquire student achievement gains, they may be more susceptible to lower evaluation scores, particularly if years of experience are not taken into consideration. Hence, what was previously described by Ingersoll (2003) as a "revolving" door of teacher turnover, high-stakes teacher evaluation may intensify this naturally-occurring phenomenon for beginning teachers; just as soon as teachers enter the profession, they may find themselves exiting.

THE RIPPLE EFFECT ON TEACHER EDUCATION

Unfortunately, all of these potential negative consequences also subsequently may make teaching a less attractive field. Future employees do consider performance evaluation models when conceptualizing ideas about whether or not they want to work in a particular field or company (Breaugh & Starke, 2000; Cable & Judge, 1996). Further, individuals choose jobs with value aspects aligned with their own values (Judge & Bretz, 1992). The research on preservice teachers indicates that individuals identified as "highly engaged persisters" (those that indicate high and stable levels of job satisfaction), report a desire to work with children and adolescents as a major motivating factor to enter and remain in teaching (Watt & Richardson, 2008). Teachers do not, however, list high student achievement test scores as a reason to enter the profession. If a teacher's job is determined by this factor and it is determined to be too stressful or risky, it may deter individuals from considering teaching as a future profession. Both of these factors may result in differential ability for schools and districts to recruit teachers. This may affect the teacher labor market (e.g. more teachers leave teaching and leave teaching for good) and the ripple effect will include teacher education (e.g., fewer students considering teaching as a potential career path).

CONCLUSION

Many would agree that ineffective teachers, after given ample time to improve, should be dismissed. However, it remains unclear whether or not high-stakes teacher evaluation will meet the intended outcomes of a more effective teacher workforce and improved student achievement. Lessons from past educational reforms and business employee evaluation models paint a grim future for high-stakes teacher evaluation. Furthermore, high-stakes teacher evaluation, in order to be successful, must, at a minimum, be based on highly reliable and valid measures. Highly effective teachers must be retained and they must maintain their effectiveness over time. More effective teachers must replace ineffective teachers who are dismissed. Even if these basic requirements are met, it is still questionable if high-stakes teacher evaluation will achieve the intended outcomes, and if so, if significant and practically meaningful gains will be made and, more importantly, at what cost? The possible unintended outcomes could undermine the very goal of high-stakes teacher evaluation and hinder, rather than support, student achievement. With teacher job satisfaction at an all time low (MetLife, 2012),[3] the damage to teacher morale and job satisfaction could be crippling. An increased loss of highly effective teachers would provide additional challenges to hard-to-staff schools, increasing the gap in opportunity to learn for students attending the most and least affluent schools.

The more high stakes teacher evaluation becomes and the closer such evaluation plans resemble true forced-distribution models, the greater the potential for unintended consequences. Further, the costs described above to students, teachers, schools, and teacher education do not include the financial costs of replacing dismissed teachers. This could run districts as much as $17,872 per teacher (National Commission on Teaching and America's Future [NCTAF], 2007). In addition, financial costs, and loss of time and resources could be accrued from lawsuits over unjust evaluation procedures (see Amrein-Beardsley & Collins, 2012).

Preliminary results from Race to the Top winners, such as the District of Columbia and Tennessee, should serve as important learning opportunities. As states swiftly adopt high-stakes teacher evaluation policies, we should collect data, research, correlation assessments between evaluation measures (e.g., variation across teachers, schools, and districts), information about teachers who are dismissed (e.g., years of experience), information about the teachers who replaced dismissed teachers (e.g., years of experience, effectiveness [if available]), measures of teacher job satisfaction and morale, student learning outcomes across time, and voluntary turnover, to name but a few.

Before policymakers continue forward with high-stakes teacher evaluation, it is important to revisit the ways in which financial resources, time, and energy are best spent, particularly if the basic requirements for meeting the intended outcomes cannot be accomplished. If such models cannot put in place protective mechanisms to eliminate or reduce unintended consequences, teacher evaluations should not be high stakes. The ability of policymakers to use the existing research is vital for informing an appropriate trajectory for high-stakes teacher evaluation. Furthermore, concerns raised above illustrate the ways in which high-stakes teacher evaluation, if modifications are not made, will change the landscape of American education for years to come.

Acknowledgments

I thank Tom Good and Ron Marx for their thoughtful and helpful feedback on earlier versions of this article. I also thank Candyce Jupiter for her help with manuscript preparation.

Notes

1. Under IMPACT, there are 20 different guidebooks describing the evaluation procedures related to specific roles, responsibilities and available data. For example, there are separate guidebooks for teachers with and without individual value-added data, nongeneral education teachers, and staff members (DCPS, 2012b).
2. Note that DCPS did not indicate that any teachers fell into the category of "developing."
3. In 2011, 44% of teachers reported they were very satisfied with teaching. This is a change from a generally upward trend in satisfaction of recent years. It is also a significant drop from 59% only two years earlier in 2009.

References

Abel, M. H., & Sewell, J. (1999). Stress and burnout in rural and urban secondary school teachers. *The Journal of Educational Research, 92,* 287–293.

Alliance for Excellent Education. (2004). *Tapping the potential: Retaining and developing high-quality new teachers.* Washington, DC: Alliance for Excellent Education.

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10*(18). Retrieved from http://epaa.asu.edu/epaa/v10n18/

Amrein-Beardsley, A., & Collins, C. (2012). The SAS Education Value-Added Assessment System (SAS® EVAAS®) in the Houston Independent School District (HISD): Intended and unintended consequences. *Education Policy Analysis Archives, 20*(12). Retrieved from http://epaa.asu.edu/ojs/article/view/1096

Banchero, S., Maher, K., Lee, C. E., & Yadron, D. (2012, September 11). Chicago teachers go on strike. *Wall Street Journal (*Eastern ed.), pp. A1–A2.

Banchero, S., Porter, C., & Belkin, D. (2012, September 19). Union vote ends strike by teachers in Chicago. *Wall Street Journal* (Eastern ed.), pp. A1–A2.

Berliner, D. (2013). Exogenous variables and value-added assessments: A fatal flaw. *Teachers College Record.*

Betoret, F. D. (2006). Stressors, self-efficacy, coping resources, and burnout among secondary school teachers in Spain. *Educational Psychology, 26,* 519–539.

Bill & Melinda Gates Foundation. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Seattle, WA: Author. Retrieved from www.metproject.org/downloads /MET_Gathering_Feedback_Research_Paper.pdf

Bossidy, L., & Charan, R. (2002). *Execution: The discipline of getting things done.* New York, NY: Random House.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2008). *Teacher attrition and student achievement.* NBER Working Paper 14022. Retrieved from http://www.nber.org/papers/w14022

Boyd, D., Lankford, H., Loeb, S., Ronfeldt, M., & Wyckoff, J. (2011). The role of teacher quality in retention and hiring: Using applications-to-transfer to uncover preferences of teachers and schools. *Journal of Policy and Management, 30*(1), 88–110

Boyle, M. (2001, May 28). Performance reviews: Perilous curves ahead. *Fortune, 143,* 187.

Breaugh, J. A., & Starke, M. (2000). Research on employee recruitment: So many studies, so many remaining questions. *Journal of Management, 26,* 405–434. doi:10.1177/014920630002600303.

Bretz, R. D. J., Milkovich, G. T., & Read, W. (1992). The current state of performance appraisal research and practice: Concerns, directions, and implications. *Journal of Management, 18,* 321–352.

Cable, D., & Judge, T. A. (1996). Person-organization fit, job choice decisions, and organizational entry. *Organizational Behavior and Human Decision Processes, 67,* 294–311.

Callahan, R. E. (1962). *Education and the cult of efficiency: A study of the social forces that have shaped the administration of the public schools.* Chicago, IL: The University of Chicago Press.

Campbell, D. T. (1976). *Assessing the impact of social change* (Technical report). Hanover, NH: Dartmouth College.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood (NBER Working Paper No. 17669). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w17699

Clark K. (2003, January 13). Judgment day. *U.S. News and World Report, 134*, 31–32.

Cloudt, C., & Stevens, N. (1995). *Texas teacher retention, mobility, and attrition* (Policy Research Report). Austin, TX: Texas Education Agency.

Cohen, S., Rothbart, M., & Phillips, S. (1976). Locus of control and the generality of learned helplessness in humans. *Journal of Personality and Social Psychology*, *34*(6), 1049–1056. doi:10.1037/0022-3514.34.6.1049

Collins, C., & Amrein-Beardsley, A. (2013). Putting growth and value-added models on the map: A national overview. *Teachers College Record.*

Crothers, L. M., Kanyongo, G. Y., Kolbert, J. B., Lipinski, J., Kachmar, S. P., & Koch, G. (2011). Job stress and locus of control in teachers: Comparisons between samples from the USA and Zimbabwe. *International Review of Education, 56,* 651-669.

Davey, M. (2012, September 10). Teachers' strike in Chicago tests mayor and union. *New York Times.* Retrieved from http://www.nytimes.com/2012/09/11/education/teacher-strike-begins-in-chicago-amid-signs-that-deal-isnt-close.html?pagewanted=all&_r=0

Darling-Hammond, L. (1999). *Teacher quality and student achievement: A review of state policy evidence.* Seattle, WA: University of Washington, Center for the Study of Teaching and Policy.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan, 93*(6), 8–15.

Darling-Hammond, L. & Haertel, E. (2012, November 5). A better way to grade teachers. *Los Angeles Times.* Retrieved from http://articles.latimes.com/2012/nov/05/opinion/la-oe-darling-teacher-evaluations-20121105

Dillion, S. (2010, December 7). Top test scores from Shanghai stun educators. *New York Times.* Retrieved from http://www.nytimes.com/2010/12/07/education/07education.html?pagewanted=all

District of Columbia Public Schools. (2011a). *An overview of IMPACT* [Web page]. Retrieved from http://www.dc.gov/DCPS/In+the+Classroom+Ensuring+Teacher+Success/IMPACT+(Performance+Assessment)/An+Overview+of+IMPACT

District of Columbia Public Schools. (2011b*). IMPACTplus* [Web page]. Retrieved from http://www.dc.gov/DCPS/In+the+Classroom/Ensuring+Teacher+Success/IMPACT+%28Performance+Assessment%29/IMPACTplus

District of Columbia Public Schools. (2011c). *DCPS continues to strengthen workforce* [Web page]. Retrieved from http://dc.gov/DCPS/About+DCPS/Press+Releases+and+Announcements/Press+Releases/DCPS+Continues+to+Strengthen+Workforce

District of Columbia Public Schools. (2012a). *General education teachers with individual value-added student achievement data.* Washington, DC: Author. Retrieved from http://www.dc.gov/DCPS/Files/downloads/In-the-Classroom/IMPACT%20Guidebooks/IMPACT-2012-Grp1%20Aug27.pdf

District of Columbia Public Schools. (2012b). IMPACT guidebooks 2012–2013. Retrieved from http://www.dc.gov/DCPS/Files/downloads/In-the-Classroom/IMPACT%20Guidebooks/IMPACT-2012-Grp2%20Aug27.pdf

District of Columbia Public Schools. (2012c). *General education teachers without individual value-added student achievement data.* Washington, DC: Author. Retrieved from http://www.dc.gov/DCPS/Files/downloads/In-the-Classroom/IMPACT%20Guidebooks/IMPACT-2012-Grp2 Aug27.pdf

Fang, Z. (1996). A review of teacher beliefs and practices. *Educational Research, 38,* 47–65.

Folger, R., & Konovsky, M. A. (1989). Effects of procedural and distributive justice on reactions to pay raise decisions. *Academy of Management Journal, 32,* 115–130. doi:10.2307/256422.

Folts, J. D. (1996). *History of the University of the State of New York and the state education department, 1784–1996.* Albany, NY:

Author.

Foster, K. (2012, April 12). Report warns US educational failures pose national security threat. *FoxNews.com*. Retrieved from http://www.foxnews.com/us/2012/04/12/report-warns-us-educational-failures-pose-national-security-threat/

Goldhaber, D., Gross, B., & Player, D. (2007). *Are public schools really losing their "best"? Assessing the career transitions of teachers and their implication for the quality of the teacher workforce* (Working Paper 12). Washington, DC: Center for Analysis of Longitudinal Data in Education Research, Urban Institute.

Good, T. (2013). What do we know about how teachers influence student performance on standardized tests, and why do we know so little about other outcomes? *Teachers College Record.*

Good, T. L., Biddle, B. J., & Brophy, J. E. *(*1975*). Teachers make a difference*. New York, NY: Holt, Rinehart & Winston.

Grote, D. (2005). *Forced ranking: Making performance management work*. Boston, MA: Harvard Business School Press.

Guin, K. (2004). Chronic teacher turnover in urban elementary schools. *Education Policy Analysis Archives, 12*(42). Retrieved from http://epaa.asu.edu/epaa/v12v42/

Hall, P. M. (1983). A social construction of reality. *The Elementary School Journal, 84*(2), 142–148.

Hanushek, E. A., Kain, J. F., O'Brien, D. M., & Rivkin, S. G. (2005, February). *The market for teacher quality* (Working Paper 11154). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w11154

Hanushek, E. A., & Rivkin, S. G. (2010). *Constrained job matching: Does teacher job search harm disadvantaged urban schools?* (Working Paper 15816). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers /w15816

Harris, D. N. (2011). *Value-added measures in education: What every educator needs to know.* Cambridge, MA: Harvard Education Press.

Herlihy, C., Karger, E., Pollard, C., Hill, H. C., Kraft, M. A., Williams, M., & Howard, S. (2013). State and local efforts to investigate the validity and reliability of scores form teacher evaluation systems. *Teachers College Record.*

Huston, J. (1989). Teacher burnout and effectiveness: A case study. *Education, 110,* 70–78.

Ingersoll, R. (2001). Teacher turnover and teacher shortages: An organizational analysis. *American Educational Research Journal, 38*(3), 499–534.

Ingersoll, R. M. (2003). *Is there really a teacher shortage?* Seattle, WA: University of Washington, Center for the Study of Teaching and Policy.

Ingersoll, R., Merrill, L., & Consortium for Policy Research in Education. (2012). *Seven trends: The transformation of the teaching force*. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, Canada.

Ingersoll, R., & Perda, D. (2012). *How high is teacher turnover and is it a problem?* Philadelphia, PA: Consortium for Policy Research in Education, University of Pennsylvania.

Jenkins, H. W. (2001, July 18). How to execute 10%, nicely. *Wall Street Journal,* p. A19.

Johnson, S., Berg, J., & Donaldson, M. (2005). *Who stays in teaching and why: A review of the literature on teacher retention.* The Project on the Next Generation of Teachers. Cambridge, MA: Harvard Graduate School of Education.

Jones, M. G., Jones, B. D., & Hargrove, T. (2003). *The unintended consequences of high-stakes testing*. Lanham, MD: Rowman & Littlefield.

Judge, T. A., & Bretz, R. D. (1992). Effects of work values on job choice decisions. *The Journal of Applied Psychology, 77*, 261–271. doi:10.1037/0021-9010.77.3.261

Kenny, D. (2012, October 14). Want to ruin teaching? Give ratings [Editorial]. *The New York Times: The Opinion Pages*. Retrieved from http://www.nytimes.com/2012/10/15/opinion/want-to-ruin-teaching-give-ratings.html?_r=1&

Kersting, N. B., Chen, M., & Stigler, J. W. (2012). Value-added teacher estimates as part of teacher evaluations: Exploring the effects of data and model specifications on the stability of teacher value-added scores. *Educational Policy Analysis Archives, 21*(7). Retrieved from http://epaa.asu.edu/ojs/article/view/1167

Kokkinos, C. M. (2007). Job stressors, personality and burnout in primary school teachers. *British Journal of Educational Psychology, 77,* 229–243.

Kyriacou, C. (2001). Teacher stress: Directions for future research. *Educational Review, 53*, 27–35.

Kwoh, L. (2012, January 31). 'Rank and yank' retains vocal fans. *The Wall Street Journal*. Retrieved from http://online.wsj.com/article/SB10001424052970203363504577186970064375222.html?ru=yahoo&mod=yahoo_hs

Lavigne, A. L., & Good, T. L. (2013). *Teacher and student evaluation: Moving beyond the failure of school reform*. New York, NY: Routledge.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44*(1), 47–67.

Lopez, R. (2010, September 15). Union stages protest against Times. *Los Angeles Times*. Retrieved from http://www.latimes.com/news/local/teachers-investigation/la-me-teachers-union-protest,0,4982349.story

Lunenberg, F., & Cadavid, V. (1992). Locus of control, pupil control ideology, and dimensions of teacher burnout. *Journal of Instructional Psychology, 19,* 13–22.

McBriarty, M. A. (1988). Performance appraisal: Some unintended consequences. *Public Personnel Management, 17,* 421–434.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67–101.

Meisler, A. (2003). Dead man's curve. *Workforce, 82,* 44–49.

MetLife, Inc. (2012). *The MetLife Survey of the American Teacher: Teachers, parents, and the economy*. New York, NY: Author. Retrieved from https://www.metlife.com/assets/cao/contributions/foundation/american-teacher/MetLifeTeacher-Survey-2011.pdf

Murnane, R. J. (1984). Selection and survival in the teacher labor market. *The Review of Economics and Statistics, 66*(3), 513–518.

Murphy, K. R., & Cleveland, J. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Thousand Oaks, CA: Sage.

National Council on Teacher Quality. (2011). *State of the states: Trends and early lessons on teacher evaluation and effectiveness policies*. Washington, DC: Author. Retrieved from http://www.nctq.org/p/publications/docs/nctq_stateOfTheStates.pdf

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Department of Education.

National Commission on Teaching and America's Future (NCTAF). (2007). *The high cost of teacher turnover*. Washington, DC: Author.

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

Orfield, G., & Kornhaber, M. L. (Eds.). (2001). *Raising standards or raising barriers? Inequality and high-stakes testing in public education*. New York, NY: The Century Foundation Press.

Patten, S. (1911). An economic measure of school efficiency. *Educational Review, 41,* 467–469.

Payne, C. M. (2008). *So much reform, so little change: The persistence of failure in urban schools*. Boston, MA: Harvard Education Press.

Pigge, F. L., & Marso, R. N. (1997). A seven year longitudinal multi-factor assessment of teaching concerns development through preparation and early years of teaching. *Teaching and Teacher Education, 13*(2), 225–235. doi:10.1016/S0742-051X(96)00014-5

Pfeffer, J., & Sutton, R. I. (2006). Evidence-based management. *Harvard Business Review, 84,* 62–74.

Ravitch, D. (2010). *The death and life of the great American school system: How testing and choice are undermining education*. New York, NY: Basic Books.

Rice, J. M. (1893). *The public-school system of the United States*. New York, NY: The Century Co.

Rice, J. M. (1913). *Scientific management in education*. New York, NY: Hinds, Noble, & Eldredge.

Roch, S. G., Sternburgh, A. M., & Caputo, P. M. (2007). Absolute vs. relative performance rating formats: Implications for fairness and organizational justice. *International Journal of Selection and Assessment, 15,* 302–316. doi:10.1111/j.1468-2389.2007.00390.x.

Ronfeldt, M., Lankford, H., Loeb, S., & Wyckoff, J. (2011, June). *How teacher turnover harms student achievement* (Working Paper

17176). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w17176

Rynes, S., Gerhart, B., & Parks, L. (2005). Personnel psychology: Performance evaluation and pay for performance. *Annual Review of Psychology, 56,* 571–600.

Sadowski, C., & Blackwell, M. (1985). Locus of control and perceived stress among student-teachers. *Psychological Reports, 56,* 723–726.

Sadowski, C., & Blackwell, M. (1987). The relationship of locus of control to anxiety among student teachers. *College Student Journal, 21,* 187-189.

Sanders, W. L., & Horn, S. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*(3), 247–256.

Sawchuk, S. (2013, February 5). Teachers' ratings still high despite new measures. *Education Week*. Advance online publication. Retrieved from http://www.edweek.org/ew/articles/2013/02 /06/20evaluate_ep.h32.html?tkn=ZSTF0t1L9D3DuLKhRwCfNuY2HyYg2z2zSJUw&cmp=ENL-EU-NEWS1

Schwarzer, R., & Hallum, S. (2008). Perceived teacher self-efficacy as a predictor of job stress and burnout: Mediation analysis. *Applied Psychology: An International Review, 57,* 152–171.

Scullen, S. E., Bergey, P. K., & Aiman-Smith, L. (2005). Forced distribution rating systems and the improvement of workforce potential: A baseline simulation. *Personnel Psychology 58,* 1–32.

Shen, J. (1991). Teacher retention and attrition in public schools: Evidence from SASS91. *The Journal of Educational Research, 91*(2), 81–89.

Skaalvik, E. M., & Skaalvik, S. (2007). Dimensions of teacher self-efficacy and relations with strain factors, perceived collective teacher efficacy, and teacher burnout. *Journal of Educational Psychology, 99,* 611–625.

Stossel. J. (2006, January 13). John Stossel's 'stupid in America' [Online article]. Retrieved from http://abcnews.go.com /2020/Stossel/story?id=1500338#.UHYATERqpMo

Strauss, V. (2012, April 20). Education reform protests pick up stream [Web log post]. *The Washington Post*. Retrieved from http://www.washingtonpost.com/blogs/answer-sheet/post/education-reform-protests-pick-up-steam/2012/04/19 /gIQA8KiXUT_blog.html

Tennessee Department of Education. (n.d.a). *Calculating the effectiveness rating*. Retrieved from http://team-tn.org/assets /educator-resources/Calculating_the_Effectiveness_Rating.pdf

Tennessee Department of Education. (n.d.b). *Tenure Q & A* [Web page]. Retrieved from http://team-tn.org/assets/educator-resources/Tenure_QandA.pdf

Tennessee Department of Education. (2012a). *Teacher evaluation in Tennessee: A report on Year 1 implementation*. Nashville, TN: Author. Retrieved from http://team-tn.org/assets/misc/Year 1 Evaluation Report TNDOE.pdf

Tennessee Department of Education. (2012b). *Acceptable range of results*. Nashville, TN: Author. Retrieved from http://www.team-tn.org/assets/misc/E- Acceptable Range_FINAL.pdf

Tennessee Department of Education. (2012c). *Number of observations*. Nashville, TN: Author. Retrieved from http://team-tn.org /assets/misc/A%20-%20Number%20of%20Observations_12_13_updated.pdf

Theobald, N., & Michael, R. (2001). *Teacher turnover in the Midwest: Who stays, leaves, and moves? Policy Issues*. Naperville, IL: North Central Regional Educational Lab.

The New Teacher Project. (2012). *The irreplaceables: Understanding the real retention crisis in urban schools*. Washington, DC: Author.

Tichy, N. M., & Sherman, S. (2001). *Control your destiny or someone else will: Lessons in mastering change-from the principles Jack Welch is using to revolutionize GE*. New York, NY: HarperCollins.

U.S. Department of Education. (2010). *Race to the Top: Phase 2 application guidelines* [Web page]. Retrieved from http://www2.ed.gov/programs/racetothetop/applicant.html

U.S. Department of Education. (2012). *Race to the Top Fund* [Web page]. Retrieved from http://www2.ed.gov/programs /racetothetop/index.html

Valenzuela, A. (Ed.). (2005). *Leaving children behind: How "Texas-style" accountability fails Latino youth*. Albany, NY: State

University of New York Press.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*, 557–574.

Watt, H. M. G., & Richardson, P. W. (2008). Motivations, perceptions, and aspirations concerning teaching as a career for different types of beginning teachers. *Learning and Instruction, 18*, 408–428.

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review, 92,* 548–573.

Welch, J. (2001). *Jack: Straight from the gut*. New York: Warner Business Books.