Teacher evaluation practices have recently shifted due to concerns about the quality of America's public school teachers. Federal financial incentive programs such as Race to the Top (RttT), initiated in 2011, and the Teacher Incentive Fund (TIF) grants program have provoked systematic changes by incentivizing states, and thus school districts, to develop methods for identifying, and in some cases firing, America's purportedly subpar teachers.[1] Accordingly, for the first time in history, many states, districts, and administrators, are now required to evaluate teachers by methods that are up to 50% based on their "value-added," as demonstrated at the classroom-level by growth on student achievement data over time (RttT, 2011).

Though bipartisan policymakers are in many ways supportive of such increased accountability initiatives, the issue has not gone undisputed. Proponents contend that value-added methods of measuring teacher quality are not only appropriate, but also necessary for the sake of students and taxpayers. In his 2012 State of the Union Address, President Obama cited a Chetty, Friedman, and Rockoff (2011) study that found an effective teacher could raise the lifetime earnings of a student by more than $250,000 (The White House, 2012). Others have argued that firing the bottom five to eight percent of teachers and replacing them with average teachers could result in an economic growth of trillions of dollars to the U.S. gross domestic product (Hanushek, 2011).

Counter to these claims, opponents, including teachers, educational researchers, and grassroots education advocates, have responded in public and academic ways. For example, teacher evaluations were at the forefront of the 2012 Chicago Teachers Strike due to the heavy reliance evaluations were to have on student achievement data (Tareen, 2012). Diane Ravitch, an education scholar and blogger about educational issues, has devoted nearly 500 posts on the topic of teacher evaluations alone.[2] Additionally, critiques of the Chetty et al. (2011) study indicated that increased earning potential resulting from effective teachers broke down to less than $20 per week per student (Baker, 2012), that the study was based on data prior to NCLB (Winerip, 2012), and that the researchers contradicted themselves with their findings, thus invalidating their claims (Adler, 2013). In all, opponents have argued that the

---

[1] The Teacher Incentive Fund (TIF) is a competitive grants program that incentivizes states to implement performance-based compensation systems for teachers and principals in high-needs schools. The compensation systems must be based at least in part on student achievement gains. Four cohorts of recipients have been awarded funds since 2006. For more information, visit: http://www2.ed.gov/programs/teacherincentive/index.html.

[2] See http://dianeravitch.net/category/teacher-evaluations/

current methods of measuring teacher effectiveness based on student growth are vastly flawed, primarily in terms of reliability, validity, bias, and fairness (Amrein-Beardsley, 2008; Baker, Oluwole, & Green, 2013; Berliner, 2013; Hill, Kapitula, & Umland, 2011; Papay, 2010).

The debate has done little to slow the momentum of policy implementation, as 44 states and the District of Columbia have thus far passed policies or legislation requiring the use of student growth data in their teacher evaluation systems (Collins & Amrein-Beardsley, 2014). Consequently, the almost three million teachers in America's public schools are in some way impacted by these policies. Teachers experience the effects of this to varying degrees depending on the policies in the state or district in which they teach. For example, some teachers' salaries and/or bonuses are based on their value-added scores and/or teacher evaluations, and some teachers can be fired for low scores.

To better understand the implications of VAM-based teacher evaluation policies, the authors of this paper first sought to understand the historical and socially-situated problem behind such practices. Then, the authors explored the current state of VAM-based procedures to depict the ways in which the policy is being realized in practice. As such, the authors conducted a review of literature, reports, and U.S. education policies to examine this controversial topic of teacher evaluation that continues to sweep the nation.

Theoretical Perspective

In 1995, Berliner and Biddle published the book, *The Manufactured Crisis*, in which they used data that they collected and analyzed to show that the increasingly fearful U.S. public had been misled by policymakers and school reform enthusiasts. They homed in on *A Nation at Risk* (1983) as being one of the most damaging and grossly exaggerated reports to affect society's view of public education. Berliner and Biddle argued that failing schools were not a product of poor quality schools and teachers, but were instead a product of a much more convoluted issue—poverty.

Socioeconomic status (SES) has long been accepted as a significant, if not the most significant, factor of student achievement in terms of large-scale standardized achievement tests (Anyon, 2005; Berliner, 2006, 2013; Biddle, 2001; Rothstein, 2004). [3] As such, reform efforts, often under the guise of

[3] The authors acknowledge that in the U.S., socioeconomic status is inextricably connected to race, and thus has implications for how we view educational equity, as well as the evaluations of students, teachers, schools, etc. Though a separate treatment of race is beyond the scope of this paper, the authors would like to refer interested readers to Johnson (2011) for a compelling analysis of the effects of

"accountability" and "standard-based reform," that seek to measure and oftentimes punish students, teachers, principals, and schools for such achievement scores, are simplistic in that they misdiagnosis the underlying issue of "failing" schools. The reality is that by the time students turn 18, about 90 percent of their lives have been spent with their family, in their neighborhood, and not inside a school or classroom (Berliner, 2012). These out-of-school factors (e.g., poverty, home life, health) are up to three times more powerful than school and teacher factors (Berliner, 2009), which would likely outweigh the effects of even the most talented teachers. Based largely on student achievement scores, the recent accountability emphasis on teacher evaluations ignores such realities and instead blames teachers for what is beyond their control.

Teacher Evaluation Practices as Policy

A policy (formal or informal) in any given context serves the distinct purpose of solving some problem. Banks establish lending policies to reduce risk of defaults. City officials establish local traffic laws to prevent automobile accidents. Given the function of a policy to solve (or prevent) a problem, there also exists a tethered consequence of policy that must be carefully regarded that, by its very design, a solution to a problem inevitably limits the scope of the problem, which limits the ability to address and/or recognize other potential causes of the problem (Goodwin, 1996). In other words, a policy is designed to target the cause of a given problem and (hopefully) fix it, thus solving the problem. However, if the policy is aimed at the wrong cause, then not only does the problem remain unsolved, but the root causes of the problem often go unexplored. Additionally, unintended consequences might ensue, further complicating the problem at hand.

In order to think about policy in terms of problem and solution, one must attempt to understand (1) the genesis of the problem that is meant to be solved by the policy, (2) the way in which the policy addresses the problem, and (3) the outcomes of the policy. As such, policy can be conceptualized as a three-tier framework that ranges from abstract societal ideas about a social system (e.g., education, healthcare, public safety) to concrete practices that occur between people within a social system (Burch, 2007).

In terms of teacher evaluation practices, the three-tier framework can serve as a lens through which to analyze the current state of teacher evaluation policies that are based in significant part on student growth data (see Table 1).

---

segregation and desegregation on students' long-term attainments in terms of schooling, earnings, and health.

Table 1. *Three-Tier Policy Framework and Corresponding Questions*

	First Tier	Second Tier	Third Tier
Tier Descriptions	Societal beliefs shaped by cultural factors	Institutions or normative systems (e.g., policies) that concretize the ideologies of the society at large (Little, 2012)	Social actors and lived experiences in particular places and times
Questions raised	What ideological problem(s) and cause(s) is VAM-based teacher evaluation policies attempting to solve?	What and how are mechanisms are used to carry out these policies? How are teachers positioned relative to these VAMs?	What are the outcomes of VAM-based teacher evaluations?

To better understand the answers to these questions, the authors organized the following literature review within the three-tier framework as referenced in Table 1.

Tier 1: The Ideological Foundation of VAM-Based Teacher Evaluation Policies

The Soviet Union's launch of Sputnik in 1957 amplified America's fear of communism and transformed the function of the public schools to an idealized one that could reaffirm the U.S. as the global leader (Steeves, Bernhardt, Burn, & Lombard, 2009). In his 1958 State of the Union Address, President Eisenhower pointed directly at the schools as one way to combat the Soviet threat, stating, "…we have tremendous potential resources on … nonmilitary fronts to help in countering the Soviet threat: education, science, research, and, not least, the ideas and principles by which we live," (The Dwight D. Eisenhower Presidential Library, Museum, and Boyhood Home, 2012). Eisenhower's proposition and use of fear tactics paved the way for future education policy initiatives, as well as a rhetorical agenda that policymakers would continue to ensue for decades to come (Johanningmeier, 2010).

A decade later, the Civil Rights Act of 1964 required a national report on the equal educational opportunities available for all individuals, catalyzing an accountability movement in the U.S. public education system. Sociologist James Coleman (1966) found inequities across schools including class sizes, student

achievement levels, school quality, school resources, and teacher quality as measured by the education levels and training of teachers. In his influential Coleman Report, he reported that teacher quality had the greatest impact on student achievement compared to all other school-related factors. The Coleman Report first introduced the impact of school inputs on student achievement and demonstrated that variation in teacher quality had a cumulative effect on students as they progressed through school (Hanushek, 1979).

Noting the inequities highlighted by the Coleman Report, Hanushek (1971) explained that improving the equitable distribution of resources was difficult because so much remained unknown about the relationship between educational inputs (i.e., teachers, curricula, peer students, facilities) and outputs (i.e., multidimensional factors composed of students' achievement and attitudinal changes). Prior to the 1970s, societal emphasis was placed on educational inputs instead of outputs, meaning relatively little was known about how schools and teachers actually affected the education process. There had been little to no historical data available at the individual student-level on how their achievement was impacted by teachers and schools. Instead, it was assumed that tenure and advanced college education resulted in more effective teachers and increased student learning; however, no studies had yet evaluated these hypotheses (Hanushek, 1971).

To further investigate the relationship between inputs and outputs, Hanushek (1970) conducted a study in a school district in southern California where he tracked students from first through third grade to examine the relationship between school system inputs and outputs "as measured by achievement scores and attitudinal change" (Hanushek, 1970, p. IV). His model used data from each student's education level (via first grade Stanford Achievement Test scores) to determine the value-added by measuring gains in achievement during the second and third grades. Other inputs in Hanushek's model included socioeconomic status, peer classmates' influence, innate abilities (e.g., IQ scores), and school influences. These inputs were based on Hanushek's hypothesis that tenure and further schooling equated to higher quality teaching and that class assignments had a beneficial effect on education. Hanushek (1970) found that significant differences in the performance of white children were dependent on the teacher, regardless of the student's socioeconomic status. However, Hanushek was unable to identify the characteristics of effective teachers and thus continued his work by applying the economic notion of inputs and outputs in education.

With traditional input-output models in an economic or manufacturing setting, two production processes applying the same inputs should result in the

same outputs, and any differences would indicate inefficiencies.[4] In education however, students with the same inputs (e.g., school, classroom, teacher) can most certainly yield different achievement outputs, which are not necessarily issues of inefficiency, rather issues that are beyond the means of the school (i.e., home life, health, and most importantly, poverty level). Despite the inability of the input-output model to identify inefficiencies in the education process, Hanushek (1979) believed the model could be useful in providing information on characteristics of teaching that could be replicated in hopes of reaching desirable outcomes in student achievement.

Hanushek's econometric model was one of the first *value-added* models derived from conceptual needs and not based on data availability. Hanushek's model was also one of the first to include inputs with cumulative influence (e.g., family background influences, classroom or peer influence, and school influence) on student achievement, which he believed had lasting impacts on student achievement year to year (Hanushek, 1979). His foundational studies of value-added measures, particularly to measure teacher inputs, were timely as education reform at the national level was about to focus more heavily on teacher quality.

A Nation at Risk. The potential for rigorous accountability mechanisms was even more luring after the release of *A Nation at Risk* in 1983. The authors of the report lambasted the public education system and, via alleged evidence, initiated a growing fear about U.S. public schools and their ability to educate students for a global rivalry. Critics of the report warned against the National Commission on Excellence in Education's use of fear tactics and claimed that the report distorted the reality of the public education system for political motivations, which was later termed the *manufactured crisis* by Berliner and Biddle (1995). Regardless, public officials espoused the ideas of the report, subsequently transforming the ways in which people thought about and acted upon student achievement, evaluation, accountability, and teacher effectiveness (Johanningmeier, 2010; Koretz, 1996). A new level of expectations for public education had emerged, positioning schools and teachers as the exclusive way of saving students from global defeat, or conversely, as the ones who can detrimentally deter future success. This marked what would become a nation obsessed with testing, evaluating, and accountability, and thus, accountability policies.

The explicit policy impact of *A Nation at Risk* was first realized in the 1990s with the reauthorization of the Elementary and Secondary Education Act

---

[4] While the discussion of whether education could and should be commodified into inputs/outputs is a valid one, it is not the focus of this paper, and Hanushek was following suit of others during this era, by applying econometric models to education.

(ESEA) and the Goals 2000: Educate America Act, which established a standards-based education model (Schwartz & Robinson, 2000). The next reauthorization of ESEA in 2002 established the No Child Left Behind (NCLB) Act, which introduced a new framework for accountability in which students, schools, and districts were required to meet state-developed standards as measured by state-developed assessments. Failure to meet such standards resulted in harsh, but intended, consequences ranging from students being retained for failure to pass state tests, schools losing federal funds for not making adequate progress, and districts being taken over by the state for failure to meet specific goals. Not only did these intended consequences restructure the education system, but the unintended consequences, such as narrowed curriculum, teaching to the test, and excessive testing, led to a massive pushback from educators and educational researchers (Amrein & Berliner, 2002; Darling-Hammond, 2007; 2010; Johnson & Johnson, 2005; Menken, 2006; Ravitch, 2010; Smyth, 2008).

After more than a decade of attempting to reach the ultimate goal of NCLB—that every student in the country be "proficient" in reading/language arts and mathematics by the year 2014—the U.S. Secretary of Education, Arne Duncan, reported that approximately 82% of schools were likely to fail to meet this goal (U.S. Department of Education, 2011). Thus, instead of forcing states to accept the consequences that had been planned and that the government was likely incapable of enforcing with such a large number of schools, Secretary Duncan presented states with a way out. Little was it realized, however, that the "way out" included plans for evaluating schools and teachers that were even more reliant on student test scores and perhaps in a more misguided way than NCLB.

Race to the Top. Simultaneously, The New Teacher Project released a report called "The Widget Effect," purporting that, once again, America's public school children were in danger (Weisberg, Sexton, Mulhern & Keeling, 2009); this time faulty teacher evaluations were to blame for U.S. student achievement lagging behind in the global economy.[5] The authors condemned school administrators' inability to distinguish good teachers from bad, while likening teachers to "widgets," or simply "interchangeable parts," (Weisburg, Sexton, Mulhurn, & Keeling, 2009, p. 4). They blamed inadequate teacher evaluation systems, which by their claims rated, on average, 99% of all teachers as effective and 1% the inverse (Weisberg et al., 2009). It seemed the

[5] The New Teacher Project is an organization that helps "districts recruit, certify and hire great teachers – those who not only show promise, but who also demonstrate a track record of raising student achievement," (TNTP.org, 2014). It began under the former Washington D.C. Public Schools Chancellor, Michelle Rhee, who is known for her efforts in attaching student achievement scores to teacher evaluations.

country faced yet another "manufactured crisis" (Berliner & Biddle, 1995), but akin to the influence of *A Nation at Risk*, this new report coupled with similar studies, had significant political influence (Corcoran, 2010; Goldhaber & Hansen, 2010; Hanushek, 2011). Thus, the race was on for a more objective, discerning teacher evaluation system that could "properly" identify effective, average, and ineffective teachers.

RttT (2011) and other post-NCLB policy initiatives, such as the aforementioned TIF grants program, adopted the "widget effect" ideology that schools were failing, teachers were to blame, and that by holding teachers accountable (i.e., punishing bad teachers and rewarding good teachers), teachers would work harder and teach better. Popular media sources, including, for example, news journalists, documentarians, and film producers who had subscribed and/or contributed to these propaganda, helped disseminate, reaffirm, and perpetuate these ideological perspectives in the greater public domain by means of emotive petitions and appeals. For example, some filmmakers used full-length movies, such as *Waiting for Superman* and *Won't Back Down*, to depict teachers and teachers unions as the epitome of the education "crisis," (Dalton, 2013).

Concurrently, scholars have heavily criticized the ways in which the concept of accountability has manifested in these various educational policies (e.g., NCLB, RttT) as well as the now widespread inclusion of accountability mechanisms such as VAMs. Scholars and other critics have denounced the fundamental and often false assumptions associated with the need for such accountability mechanisms (Berliner, 2006; Rubin, Stuart, & Zanutto, 2004). Some challenge the notion that increased accountability systems based on high-stakes tests can improve educational quality and instead posit that such systems ignore and reinforce inequalities based on socioeconomic factors and race (Au, 2009; Orfield & Kornhaber, 2000). Others claim that such systems produce unintended consequences, such as schools excluding particular students from test taking by encouraging students to drop out or by re-classifying students as special education (Haney, 2000; Klein, Hamilton, McCaffrey, & Stecher, 2000). Such practices do little, if anything, to address the root problems of educational quality.

Tier 2: The Mechanisms of VAM-Based Teacher Evaluation Policies

A predominance of the VAM-based teacher evaluation literature has focused on the mechanisms, or instruments, used to carry out contemporary teacher evaluation policies. Most often explored are the methodological concerns associated with RttT-fashioned teacher evaluation systems. Researchers in this branch of the literature are most concerned with the reliability and validity of the statistical instruments, such as VAMs, intended to measure the causal

relationships between a teacher's instruction and students' learning.

*Value-added models***.** VAMs are statistical tools used to measure the purportedly causal relationship between a teacher's instruction and the respective students' learning, by measuring student growth over time on large-scale standardized achievement tests while controlling for some student characteristic variables (e.g., prior testing history and demographics) and some classroom and school level characteristic variables (e.g., class size, school demographics). VAMs are intended to objectively measure the amount of "value" that a teacher "adds" to (or detracts from) a student's learning over a school year.

Though variations of VAMs exist with different inputs or variables and controls included in the models, the output is always measured by student growth on some type of large-scaled standardized achievement test. According to Harris (2011), reliance on such tests inevitably marginalizes a majority—approximately 70%—of teachers because only teachers who teach grade levels and content areas with standardized tests (commonly fourth through eighth grades in the subjects of mathematics and reading/language arts) are typically included in the models. This inability to accurately represent the work of a great portion of teachers gets at a fundamental issue with fairness in the use of VAMs; it has led many states to attribute an aggregate, school-level value-added score to the non-tested grade level and content area teachers (Collins & Amrein-Beardsley, 2014). In other words, a majority of teachers' VAM scores are based on students and/or subjects that they do not teach. Problems with fairness also manifest in terms of the statistical concerns with the VAMs as they are currently designed and implemented.

When using measurement tools such as VAMs, interpretations and uses derived from the tools matter even more than the numbers produced (Cronbach & Meehl, 1955; Kane, 2006, 2013; Messick, 1975, 1980, 1989, 1995). In other words, in order to fully understand the outputs yielded by a statistical model, one should first understand and acknowledge the model's limitations. Yet given the increased reliance on such scores, this critical notion is seemingly ignored by VAM advocates, and the specifics regarding how value-added scores are determined is typically not available to administrators and teachers in accessible, easy-to-digest formats. The lack of information about the limitations of VAMs ultimately positions administrators and teachers as unassuming consumers.

Reliability and VAMs. In terms of VAMs, reliability refers to the likelihood of a teacher being correctly identified as either adding or detracting value from students' learning. A key marker of reliability would be the consistency of teacher-level value-added scores from one year to the next. Of primary concern here is that evidence of reliability, or stability, is weak to moderate at best, with most value-added researchers yielding time-series

correlations within the range of $0.3 \leq r \leq 0.4$ (McCaffrey, Sass, Lockwood, & Mihaly, 2009; Kane & Staiger, 2012; Lockwood & McCaffrey, 2009; Newton, Darling-Hammond, Haertel, & Thomas, 2010), while some correlations are as low as $r = 0$ (Linn & Haug, 2002) or as high as $r = 0.6$ (Kersting, Chen, & Stigler, 2013). This instability can mean one of two things—either a majority of teachers' effectiveness truly fluctuates from one year to the next, or, more likely, there is a reliability problem with the models, which results in the misclassification of teachers. The question remains, how much error is too much error, especially given the often high stakes attached to such classifications?

Validity and VAMs. Researchers have questioned the evidence of value-added models' validity as well, arguing that many model types cannot fully account for the impact of uncontrollable factors (e.g., other teachers' effects, students' peer effects, summer gains/losses, outside-of-school variable effects, missing data) on yielding valid value-added estimates from which valid inferences can be made (Amrein-Beardsley, 2008; Capitol Hill Briefing, 2011; Ishii & Rivkin, 2009; McCaffrey et al., 2004; Scherrer, 2011).

Additionally, there are issues with criterion-related evidence of validity, which refers to the extent to which value-added scores align with other evaluative measures (Bill & Melinda Gates Foundation, 2013; Papay, 2010), and construct-related evidence of validity, which refers to the extent to which value-added scores actually measure the construct of interest, teaching effectiveness (Capitol Hill Briefing, 2011; Newton et al., 2010; Rothstein, 2009; 2010). First, there is a lack of statistical correlation between value-added estimates and other indicators of teacher quality, such as principal observations or teaching awards (Amrein-Beardsley & Collins, 2012; Collins, 2012). There is also a misalignment between value-added estimates derived from different tests meant to measure the same thing and administered at the same time. This misalignment is approximately $0.37 \leq r \geq 0.5$ for reading/language arts and $0.22 \leq r \geq .59$ for mathematics (Bill & Melinda Gates Foundation, 2010; Corcoran, Jennings, & Beveridge, 2011). There are also concerns when comparing estimates derived from criterion-referenced assessments to norm-referenced assessments, meaning the scores serve different purposes and do not fairly lend to comparison (Amrein-Beardsley & Collins, 2012).[6]

Bias and VAMs. Yet another point of contention with VAMs is bias (Hill et al., 2011; Newton et al., 2010; Rothstein, 2009), or the extent to which exogenous variables influence teachers' value-added scores and/or their capacities to demonstrate growth (Linn & Haug, 2002; Wright, Horn, & Sanders, 1997). For

[6] Criterion-referenced tests are designed to distinguish what students have learned or mastered. Norm-referenced tests are commonly designed to note differences in achievement among and between students in a relative manner.

example, teachers of students who typically score in the 99th percentile have a difficult time demonstrating growth because there is no room to grow – a phenomenon sometimes called the ceiling effect. While the most statistically sound method of reducing bias in these estimates might be to randomly assign students and teachers to classrooms (Raudenbush, 2004), this randomization is highly unlikely because principals find value in placing students with teachers based on students' needs (Bill & Melinda Gates Foundation, 2013; Paufler & Amrein-Beardsley, 2013).

Tier 3: The Outcomes Associated with VAM-Based Teacher Evaluation Policies

Despite the growing body of literature about the methodological issues with VAM-based teacher evaluation practices and policies, we still know very little about how the features of these teacher quality and accountability measures are understood and experienced by teachers and their evaluators in practice. Most of the existing studies, rather, have maintained a level of distance between not only the researcher(s) and their subjects (i.e., teachers), but also between the mechanisms associated with the evaluation systems/policies and the same subjects. In other words, while researchers have conducted studies to statistically test the levels of reliability and validity and the evidence of bias surrounding VAMs, very few researchers have actually asked teachers and their evaluators to report on their experiences about VAM-use in practice. This type of research would help us understand whether policies are being realized as intended. One model of such research is the Collins (2012) study of a group of teachers who are currently evaluated under a VAM-based system with high-stakes consequences (e.g., merit pay, termination).

Collins (2012) sought the perspectives of the teachers via survey methods and found that teachers reported concerns with the reliability, validity, and bias of the VAM-use in their district. Additionally, the study revealed many unintended consequences associated with the high-stakes use of the VAM, in which teachers admitted to teaching to the test, targeting instruction to students most likely to show growth, and unwillingness to collaborate or share best practices with other teachers who were seen as competitors. While the unintended consequences were troublesome, equally as troublesome was that teachers also reported little to no use of VAM scores for making instructional decisions, thus raising the question whether the undergirding of VAM-based policies is to improve existing teacher quality or simply remove teachers from the profession. Assuming the former, teachers in the Collins (2012) study overwhelmingly stated that VAM reports were vague and unclear, and that they relied on other sources of data—not VAM data—to inform them of their teaching effectiveness.

While it might be too soon to expect more empirical work on the outcomes of VAM-based teacher evaluation policies, there have been legal cases

that have resulted from questionable evaluation practices. For example, a group of Florida teachers filed a lawsuit in April, 2013 on the grounds of being evaluated based on students whom they do not teach (Jordan, 2013). Similar cases are likely to arise, as well as others due to the problems of reliability and validity with the VAMs that are currently used in state and district evaluation policies (Baker et al., 2013).

Given the early stages of the policy implementation process, there is still little known about the outcomes of VAM-based teacher evaluation policies. Future research might consider such factors as curriculum and professional development decisions based on VAM outcomes, or the effects of such systems on achievement disparities, school culture, and school finances. Another area entirely void in the current research base is what impact VAM-based policies are having on student learning and achievement scores. It is simply assumed that if teacher quality increases, student learning will simultaneously increase. Further research is certainly warranted in this area as improved test scores may actually be a result of narrowed curriculum and increased test preparation at the expense of a more well-rounded curriculum.

Discussion

The current literature on VAM-based teacher evaluation policies offers substantial grounding for answering the three questions presented in the analysis herein: (1) What is the problem that VAM-based teacher evaluation policies are meant to solve? (2) What are the mechanisms used to carry out the policies and are they appropriate? And (3) Does the policy solve the problem? Based on the literature, we have formulated answers to these three questions.

What is the problem that VAM-based teacher evaluation policies are meant to solve?

The literature suggests that the problem of a failing education system was first introduced during the Sputnik era of the 50s, reaffirmed in the 80s with the release of *A Nation at Risk*, and concretized in policy in the early 2000s with NCLB. RttT has joined its predecessors in addressing a now 60-year-old professed problem, this time directly targeting teachers as the root cause of failing schools. The main issue is that the targeted cause (e.g., poor teachers) of this problem has been supported with little (if any) empirical evidence. Therefore, suggesting that another round of increased accountability mechanisms will do anything to improve the quality of the education system is increasingly showing to have negative consequential outcomes for teachers– while even less empirical

evidence exists on how student achievement and learning outcomes have been impacted.

In fact, an overwhelming majority of the literature suggests that even though teachers are the most significant in-school factor in student achievement scores (Goldhaber, 2002; Sanders, 2000), they really only account for approximately 10-20% of student achievement score variation overall (Kennedy, 2010; Gabriel & Allington, 2011; Xu, Ozek, & Corritore, 2012), and factors such as home-life, health, poverty, etc., things well beyond the control of teachers and schools, largely influence student achievement (Berliner, 2013). This calls into question the focus of VAM-based teacher evaluation policies. Is the cause of failing schools even remotely addressed by such policies? Or is the focus on teachers limiting the scope of the problem and hindering the development of policy aimed at the more likely cause of failing schools—poverty (Anyon, 2005; Berliner, 2006; Biddle, 2001; Rothstein, 2004)? Thirty years of increased accountability policies have resulted in no evidence to suggest that more of the same will address the root causes of low student achievement scores (Au, 2009; Haney, 2000; Hursh, 2008; Klein et al., 2000; Orfield & Kornhaber, 2001).

What are the mechanisms used to carry out the policy, and are they appropriate?

First and foremost, accountability mechanisms that are designed to incentivize teachers to perform better by rewarding (e.g., merit pay) and punishing (e.g., terminating) teachers, do not work for various reasons (see, for example, the work of Ehlert, Koedel, Parsons, & Podgursky, 2012; Springer et al., 2010; Wells, 2011). They are misguided, unsophisticated, and ignore the complexities that are involved in teaching (Berliner, 2005; Gabriel & Allington, 2011; Harris, 2011; Linn, 2008; Tekwe, et al., 2004). Nonetheless, VAM (and other accountability) enthusiasts continue to advocate for increased incentive systems for teachers (Gabriel & Allington, 2011).

Even if these incentive systems were appropriate, at the heart of such systems is the VAM, which has shown to be immensely flawed in terms of reliability, validity, bias, and fairness. Such being the case, the mechanisms in place to carry out VAM-based teacher evaluation policies are far from where they should be (and likely ever will be) in order to perform the task they are currently called upon to do. Regardless, the models are accepted in their imperfect ways as a "good enough" method for teacher accountability systems (Harris, 2011).

Does the policy solve the problem?

While there is not sufficient evidence to determine whether VAM-based teacher evaluation policies are accomplishing the stated goal of increased student achievement as a result of improving teacher quality (see RttT, 2011), it is not too

soon to call upon the implications of previous policy research to make some predictions. NCLB and other accountability policies have offered little, if any, evidence to suggest that even more stringent accountability mechanisms will lead to greater student achievement. It is unlikely that VAM-based teacher evaluation policies will be any different, as unintended outcomes of the policies imply that students may actually be negatively impacted by narrowing of the curriculum and focusing instruction on those students believed to demonstrate the most growth (Amrein-Beardsley & Collins, 2012). As it stands, VAM-based policies do not and will not solve the problem of low educational quality because the policy is aimed at the wrong cause. Until the issue of poverty is more explicitly dealt with in policy, other simplistic attempts to solve the problem will continue to fail.

Conclusion

VAM-based teacher evaluation policies have taken root across the country, affecting the almost three million teachers in America's public schools, sometimes in highly consequential ways. The fundamental ideology driving teacher accountability mechanisms is grossly misguided and disproportionately focused on measuring student growth on standardized assessments. The U.S. has spent the past 30 years refining a series of accountability policies in hopes of targeting the root cause of low educational quality. This has resulted in more than 30 years of failed policy and billions of federal dollars spent, leaving little to be expected from the next attempt.

Such policies, by their very nature, have limited our scope of understanding the big picture problem masked as low educational quality. Policymakers have narrowed in so acutely on teachers, despite the limited impact that teachers ultimately have on student achievement (Kennedy, 2010; Gabriel & Allington, 2011; Xu, Ozek, & Corritore, 2012), so as to blindly ignore that which has been proven over and over again to have the most profound impact on student achievement—poverty (Anyon, 2005; Berliner, 2006; Biddle, 2001; Rothstein, 2004). The questions left to be answered now are will policymakers and educational leaders finally start using the past decades of educational research as scientific evidence? Or will our students and teachers be expected to endure another 30 years of misguided and damaging policy, which ignores the root cause of low student achievement?

Only time and more research will be able to tell us about how VAM-based teacher evaluation policies are being realized in practice. Given the widespread adoption and implementation of such policies, it is critical that we know more about the effects of such practices on a variety of education facets, such as teacher retention, educational quality, economic outcomes, student achievement, etc. More research on the outcomes of VAM-based systems will help us to understand

whether these policies are having the intended consequences as has been theorized and sold to the public.

References

Amrein, A. L. & Berliner, D. C. (2002). High-Stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10*(18), 1-74. Retrieved from http://epaa.asu.edu/epaa/v10n18/

Amrein-Beardsley, A. (2008). Methodological concerns about the Education Value-Added Assessment System (EVAAS). *Educational Researcher, 37*(2), 65-75. doi: 10.3102/0013189X08316420

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), 1-36. Retrieved from http://epaa.asu.edu/ojs/article/view/1096

Anyon, J. (2005). What "counts" as educational policy? Notes toward a new paradigm. *Harvard Educational Review, 75*(1), 65-88.

Au, W. (2009). *Unequal by design: High-stakes testing and the standardization of inequality.* New York, NY: Routledge.

Baeder, J. (2010, December 21). Gates' measures of effective teaching study: More value-added madness. *Education Week*. Retrieved from http://blogs.edweek.org/edweek/on_performance/2010/12/gates_measures_of_effective_ teaching_study_more_value-added_madness.html

Baker, B. D. (2012, January). Fire first, ask questions later? Comments on recent teacher effectiveness studies. Retrieved from http://schoolfinance101.wordpress.com/2012/01/07/fire-first-ask-questions-later-comments-on-recent-teacher-effectiveness-studies/

Baker, B. D., Oluwole, J. O., & Green, P. C. (2013). The legal consequences of mandating high stakes decisions based on low quality information: Teacher evaluation in the Race-to-the-Top era. *Education Policy Analysis Archives, 21*(5), 1-71. Retrieved from http://epaa.asu.edu/ojs/article/view/1298

Benson, B. L. (2009). Escalating the war on drugs: causes and unintended consequences. *Stanford Law & Policy Review*, *20*(2), 293.

Berliner, D. C. (2005). The place of process-product research in developing the agenda for research on teacher thinking. In P. M. Denicolo and M. Kompf (Eds.), *Teacher thinking and professional action* (pp. 325-344). New York, NY: Routledge.

Berliner, D. C. (2006). Our impoverished view of educational research. *Teachers College Record, 108*(6), 949-995.

Berliner, D. C. (2008). Research, policy and practice: The great disconnect. In Lapan, S. D. & Quartaroli (eds.), *Research essentials: An introduction to designs and practices* (pp. 295- 325). Hoboken, NJ: Jossey-Brass.

Berliner, D.C. (2009). *Poverty and potential: out-of-school factors and school success*. Boulder, CO and Tempe, AZ: Education and the Public Interest Center, University of Colorado/Education Policy Research Unit, Arizona State University. Retrieved from http://epicpolicy.org/publication/poverty-and-potential.

Berliner, D. C. (2013). Effects of inequality and poverty vs. teachers and schooling on America's youth. *Teachers College Record,* 115(12). Retrieved from: http://www.tcrecord.org/Content.asp?ContentID=16889

Berliner, D. C. (2014). Exogenous variables and value-added assessments: A fatal flaw. *Teachers College Record,* 116(1). Retrieved from: http://www.tcrecord.org/Content.asp?ContentId=17293

Berliner, D. C., & Biddle, B. J. (1995). *The manufactured crisis: Myths, fraud, and the attack on America's public schools*. Reading, MA: Addison-Wesley Publishing Company, Inc.

Biddle, B. J. (2001). *Social class, poverty, and education*. New York, NY: Routledge Falmer.

Bill & Melinda Gates Foundation. (2010, December). *Learning about teaching: Initial findings from the Measures of Effective Teaching Project*. Seattle, WA. Retrieved from http://www.gatesfoundation.org/college-ready-education/Documents/preliminary-findings-research-paper.pdf

Bill & Melinda Gates Foundation. (2013, January 8). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study.* Seattle, WA. Retrieved from http://metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf

Burch, P. (2007). Educational policy and practice from the perspective of institutional theory: Crafting a wider lens. *Educational Researcher*, *36*(2), 84-95.

Bush-Baskette, S. (2000). *The war on drugs and the black female: Testing the impact of the sentencing policies for crack cocaine on black females in the federal system.* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Database. (Order No. 9967091, Rutgers The State University of New Jersey - Newark).

Capitol Hill Briefing. (2011, September 14). *Getting teacher evaluation right: A challenge for policy makers.* A briefing by E. Haertel, J. Rothstein, A. Amrein-Beardsley, and L. Darling-Hammond. Washington DC: Dirksen Senate Office Building (research in brief). Retrieved from

http://www.aera.net/Default.aspx?id=12856

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. 17699*.) Cambridge, MA: National Bureau of Economic Research. Retrieved from http://obs.rc.fas.harvard.edu/chetty/w19423.pdf

Coleman, J. S. (1966). *Equality of educational opportunity*. U.S. Government Printing Office, Washington, D.C.: National Center for Educational Statistics.

Collins, C. (2012). Houston, we have a problem: Studying the SAS® Education Value-Added Assessment System (EVAAS®) from teachers' perspectives in the Houston Independent School District (HISD). (Doctoral dissertation). Available from Arizona State University Libraries Digital Repository. Retrieved from http://repository.asu.edu/items/16043

Collins, C. & Amrein-Beardsley, A. (2014). Putting growth and value-added models on the map: A national overview. *Teachers College Record* 116(1). Retrieved from: http://www.tcrecord.org/Content.asp?ContentId=17291

Corcoran, S. P. (2010). *Can teachers be evaluated by their students' test scores? Should they be? The use of value-added measures of teacher effectiveness in policy and practice*. Providence, RI: Annenberg Institute for School Reform. Retrieved from http://annenberginstitute.org/publication/can-teachers-be-evaluated-their-students%E2%80%99-test-scores-should-they-be-use-value-added-mea

Corcoran, S. P., Jennings, J. L., & Beveridge, A. A. (2011). *Teacher effectiveness on high- and low-stakes tests.* Manuscript submitted for publication. Retrieved from https://files.nyu.edu/sc129/public/papers/corcoran_jennings_beveridge_2011_wkg_teacher_effects.pdf

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302. doi:10.1037/h0040957

Dalton, M. (2013). How media and film portray teachers and school reform. Paper presented at the *America Educational Research Association* Annual Meeting. San Francisco, CA.

Darling-Hammond, L. (2007). Race, inequality and educational accountability: The irony of "No Child Left Behind". *Race, Ethnicity and Education, 10*(3), 245-260.

Darling-Hammond , L. (2010). *The flat world and education*. New York: Teachers College Press.

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2012, August). *Selecting growth measures for school and teacher evaluations*. Washington D. D.C: National Center for Analysis of Longitudinal Data in Education Research

(CALDER). Retrieved from www.caldercenter.org/publications/upload/WP-80.pdf

Gabriel, R. & Allington, R. (2011, April). *Teacher effectiveness research and the spectacle of effectiveness policy*. Paper Presented at Annual Conference of the American Educational Research Association (AERA), New Orleans, LA.

Goldhaber, D. (2002). What might go wrong with the accountability measures of the "no child left behind act?" *Paper presented at the meeting of Will No Child Truly Be Left Behind? The Challenges of Making this Law Work, Washington DC.*

Goldhaber, D. & Hansen, M. (2010). "Is it just a bad class? Assessing the stability of measured teacher performance." CEDR Working Paper 2010-3. Seattle, WA. Retrieved from http://www.cedr.us/publications.html

Goodwin, N. (1996). Governmentality in the Queensland Department of Education: Policies and the management of schools. *Discourse*, *17*(1), 65-74.

Haney, W. (2000). The myth of the Texas miracle in education. *Education Analysis Policy Archives, 8*(41) [On-line]. Retrieved from http://epaa.asu.edu/epaa/v8n41

Hanushek, E. A. (1970). *The value of teachers in teaching*. Santa Monica, CA: Rand Corporation. (ERIC Accession No. ED 073 089).

Hanushek, E. A. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *The American Economic Review, 61*(2) 280-288.

Hanushek, E. A. (1979). Conceptual and empirical issues in the estimation of educational production functions. *The Journal of Human Resources, 14*(3) 351-388.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review, 30(3),* 466-479.

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

Hill, H. C., Kapitula, L, & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal, 48*(3), 794-831. doi:10.3102/0002831210387916

Ishii, J., & Rivkin, S. G. (2009). Impediments to the estimation of teacher value added. *Education Finance and Policy, 4(4),* 520-536. doi:10.1162/edfp.2009.4.4.520

Johanningmeier, E. V. (2010). "A Nation at Risk" and "Sputnik": Compared and reconsidered. *American Educational History Journal, 37*(2), 347-365.

Johnson, D. D., & Johnson, B. (2005). *High stakes: Poverty, testing, and failure in American schools* (2nd Ed.). Lanham, MD: Rowman & Littlefield

Publishers.
Jordan, G. (2013, April 16). Teachers union files federal lawsuit challenging Florida teacher evaluations. *StateImpact*. Retrieved from http://stateimpact.npr.org/florida/2013/04/16/teachers-union-files-federal-lawsuit-challenging-florida-teacher-evaluations/
Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Washington, DC: The National Council on Measurement in Education & the American Council on Education.
Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. doi:10.1111/jedm.12000
Kane, T., & Staiger, D. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf
Kennedy, M. M. (2010). Attribution error and the quest for teacher quality. *Educational Researcher, 39*(8), 591-598. doi:10.3102/0013189X10390804
Kersting, N. B., Chen, M., & Stigler, J. W. (2013). Value-added added teacher estimates as part of teacher evaluations: Exploring the effects of data and model specifications on the stability of teacher value-added scores. *Education Policy Analysis Archives, 21*(7), 1-39. Retrieved from http://epaa.asu.edu/ojs/article/view/1167
Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). What do test scores in Texas tell us? *Education Policy Analysis Archives, 8*(49), 1-22. Retrieved from http://epaa.asu.edu/epaa/v8n49
Koedel, C., & Betts, J. R. (2007, April). *Re-examining the role of teacher quality in the educational production function.* Working Paper No. 2007-03. Nashville, TN: National Center on Performance Initiatives.
Koretz, D. (1996). Using student assessments for educational accountability. In E. A. Hanushek & D. W. Jorgenson (Eds.), *Improving America's schools: The role of incentives.* Washington, DC: National Academy Press.
Linn, R. L. (2008). Methodological issues in achieving school accountability. *Journal of Curriculum Studies, 40*, 699-711. doi:10.1080/00220270802105729
Linn, R L., & Haug, C. (2002). Stability of school-building accountability scores and gains. *Educational Evaluation and Policy Analysis, 24*, 29-36. doi:10.3102/01623737024001029
Little, D. (2012). Social mechanisms and meso-level causes. Paper presented at the British Society for the Philosophy of Science Annual Meeting.
Lockwood, J. R. & McCaffrey, D. F. (2009). Exploring student-teacher interactions in longitudinal achievement data. *Education Finance and*

Policy, 4(4), p. 439-467. doi:10.1162/edfp.2009.4.4.439

Lockwood, J. R., McCaffrey, D. F., & Sass, T. R. (2008). *The intertemporal stability, of teacher effect estimates.* Paper presented at the National Conference on Value-Added Modeling. Sponsored by the Wisconsin Center for Education Research (WCER), Madison, WI. Retrieved from http://www.wcer.wisc.edu/news/events/VAM%20Conference%20Final%20Papers/IntertemporalStability_McCaffreySassLockwood.pdf

Marzano, R., Livingston, D., & Frontier, T. (2011). *Effective supervision: Supporting the art and science of teaching.* Alexandria, VA:ASCD.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4*(4), 572–606. doi:10.1162/edfp.2009.4.4.572

Menken, K. (2006). Teaching to the test: How no child left behind impacts language policy, curriculum, and instruction for English language learners. *Bilingual Research Journal, 30*(2), 521-546.

Messick, S. (1975). The standard problem: meaning and values in measurement and evaluation. *American Psychologist*, 30(10), 955-66.

Messick, S. (1980 *P*). Test validity and the ethics of assessment. *American Psychologist,* 35(11), 1012-1027.

Messick, S. (1989). *Validity*. In R. L. Linn (Ed.), Educational Measurement, (3rd ed., pp. 13-103.) New York, NY: American Council on Education and Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist,* 50(9), 741-749.

Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010) Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives, 18* (23), 1-27. Retrieved from http://epaa.asu.edu/ojs/article/view/810

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.

Papay, J. P. (2010). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal,* 48(1), 163-193. doi: 10.3102/0002831210362589

Paufler, N. A. & Amrein-Beardsley, A. (2013). The random assignment of students into elementary classrooms: Implications for value-added analyses and interpretations. *American Educational Research Journal,* 20(10), 1-35.

Race to the Top (RttT) Act, Senate Bill 844 (2011). Retrieved from http://www.govtrack.us/congress/bills/112/s844

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics, 29*(1), 121-129. doi:10.3102/10769986029001121

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.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy, (4)*4, 537-571. doi:http://dx.doi.org/10.1162/edfp.2009.4.4.537

Rothstein, J. (2010, February). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, 125(1) 175-214. doi:10.1162/qjec.2010.125.1.175

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics, 29*(1), 103-116. doi:10.3102/10769986029001103

Sanders W. L. (2000). Value-added assessment from student achievement data: Opportunities and hurdles. *Journal of Personnel Evaluation in Education, 14* (4), 329-339.

Scherrer, J. (2011). Measuring teaching using value-added modeling: The imperfect panacea. *NASSP Bulletin, 95*(2), 122-140. doi:10.1177/0192636511410052

Schwartz, R. B., & Robinson, M. A. (2000). Goals 2000 and the standards movement. *Brookings Papers on Education Policy,* 173-214.

Smyth, T. S. (2008). Who is no child left behind leaving behind? *Clearing House: A Journal of Educational Strategies, Issues and Ideas, 81*(3), 133-137.

Springer, M. G., Ballou, D., Hamilton, L. S., Le, V., Lockwood, J. R., McCaffrey, D. F., Pepper, M., & Stecher, B. M. (2010). Teacher pay for performance: Experimental evidence from the Project on Incentives in Teaching (POINT). Nashville, TN: Vanderbilt University. Retrieved from http://www.rand.org/pubs/reprints/RP1416.html

Steeves, K. A., Bernhardt, P. E., Burns, J. P., & Lombard, M. K. (2009). Transforming American educational identity after sputnik. *American Educational History Journal, 36*(1), 71-87.

Tareen, S. (2012, September 13). Teacher evaluations at center of Chicago strike. *Huffington Post*. Retrieved from http://www.huffingtonpost.com/2012/09/13/teacher-evaluations- at-ce_0_n_1880264.html

Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M. E., Roth, J., …Resnick, M. B. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics, 29* (1), 11-36. doi:10.3102/10769986029001011

The Dwight D. Eisenhower Presidential Library, Museum, and Boyhood Home. (2012). *1958 State of the union address*. Retrieved from website: http://www.eisenhower.archives.gov/all_about_ike/speeches.html

The White House (2012). *Remarks by the president in the state of the union address*. Retrieved from website: http://www.whitehouse.gov/the-press- office/2012/01/24/remarks-president-state-union-address

U.S. Department of Education. (1983). *A nation at risk: The imperative for educational reform* Retrieved from http://datacenter.spps.org/uploads/SOTW_A_Nation_at_Risk_1983.pdf

U.S. Deparatment of Education (2011). *Duncan says 82 percent of America's schools could "fail" under NCLB this year*. Retrieved from website: http://www.ed.gov/news/press-releases/duncan-says-82-percent-americas-schools-could-fail-under-nclb-year

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). "The Widget Effect." *Education Digest, 75*(2), 31–35.

Wells, J. (2011, April). *Teacher responses to pay-for-performance policies: Survey results from four high-poverty, urban school districts*. Paper presented at the Annual Meeting of the American Educational Research Association (AERA), New Orleans, LA.

Winerup, M. (2012, January 15). Study on teacher value uses data from before teach-to-test era. *New York Times.* Retrieved from http://www.nytimes.com/2012/01/16/education/study-on-teacher-value-uses-data-from-before-teach-to-test-era.html?_r=0

Wright, P., Horn, S., & Sanders, W. L. (1997). Teachers and classroom heterogeneity: Their effects on educational outcomes. *Journal of Personnel Evaluation in Education, 11*(1), 57-67.

Xu, Z., Ozek, U., & Corritore, M. (2012, June). *Portability of teacher effectiveness across schools.* Washington D. C.: National Center for Analysis of Longitudinal Data in Education Research (CALDER). Retrieved from http://www.caldercenter.org/publications/upload/wp77.pdf

Authors

Jessica Holloway-Libell is a Ph.D. Candidate in the Education Policy & Evaluation program at Arizona State University. Her research interests are in education policies on teacher evaluation and value-added assessment models, as well as the discursive practices of education stakeholders as it pertains to education policies.

Clarin Collins is a graduate from the Educational Policy and Evaluation program at the Mary Lou Fulton Teachers College at Arizona State University. Her

research interests include national and local policy implementation at the classroom level, teacher influences on policymaking and implementation, and education evaluation and accountability systems. Clarin currently works as a research and evaluation officer at a private foundation in Phoenix.