Yale College
Education Studies Program

Senior Capstone Projects Student Research

Spring 2020

A Critique of Value-Added Modeling: A Mixed Methods Approach

Jack Lattimore
Yale College, New Haven Connecticut

Abstract:

This capstone project engages with Value-Added Modeling (henceforth VAM) through a mixed-methods approach. While other works on VAM tend to either follow primarily qualitative (i.e. case studies, proposals of alternative evaluation systems, analysis of political rhetoric) or quantitative (i.e. analysis of data, development of econometric methods) methodologies, I oer a mixed-methods analysis. In combining these methodologies, I endeavor to engage with VAM in a more holistic and comprehensive manner than previous literature on the subject. Specifically, I provide an analysis of the fundamental benefits and drawbacks of VAM from both qualitative and quantitative perspectives, analyze real-world case studies, and evaluate recommendations for potential alternatives. Through this wide-ranging analysis, I ground these proposals in the research literature from a variety of inter-disciplinary elds and perspectives. Furthermore, I argue that teachers can have a substantial impact on their students, and improving evaluative systems can better equip teachers to serve their students' needs.

Suggested Citation: Lattimore, J. (2020). *A Critique of Value-Added Modeling: A Mixed Methods Approach* (Unpublished Education Studies capstone). Yale University, New Haven, CT.

A Critique of Value-Added Modeling: A Mixed Methods Approach

Jack Lattimore

A Senior Capstone Project in Education Studies [1]

Yale University

May 1, 2020

[1]Many thanks to my Capstone advisors— Professors Grace Kao and Winston Lin who provided highly useful feedback and advice throughout the entire writing and editing process. I also thank Anne Mishkind and Talya Zemach-Bersin who helped me generate and refine Capstone ideas and further develop them over the course of the Capstone. Finally, I thank other members of the 2020 Education Studies cohort for the many useful peer feedback sessions on all components of the Capstone.

A Critique of Value-Added Modeling: A Mixed Methods Approach

Jack Lattimore

Abstract

This capstone project engages with Value-Added Modeling (henceforth VAM) through a mixed-methods approach. While other works on VAM tend to either follow primarily qualitative (i.e. case studies, proposals of alternative evaluation systems, analysis of political rhetoric) or quantitative (i.e. analysis of data, development of econometric methods) methodologies, I offer a mixed-methods analysis. In combining these methodologies, I endeavor to engage with VAM in a more holistic and comprehensive manner than previous literature on the subject. Specifically, I provide an analysis of the fundamental benefits and drawbacks of VAM from both qualitative and quantitative perspectives, analyze real-world case studies, and an evaluate recommendations for potential alternatives. Through this wide-ranging analysis, I ground these proposals in the research literature from a variety of inter-disciplinary fields and perspectives. Furthermore, I argue that teachers can have a substantial impact on their students, and improving evaluative systems can better equip teachers to serve their students' needs.

Keywords— Value Added Modeling, Teacher Evaluation, Mixed Methods Approaches

1 Introduction

Value Added Modeling (VAM) provides a system for evaluating teachers based on econometric models of students' standardized test score gains. Most of the current research literature on Value Added Modeling (VAM) solely emphasizes either the quantitative or qualitative component of these models. The qualitative literature places VAM within its historical context or examines its underlying political implications while the the quantitative literature tends to either derive empirical improvements to VAM estimates or provide causal evidence connecting teachers' value-added scores to students' future outcomes (e.g. salaries, academic outcomes, etc.). However, there is a lack of current research that merges these analyic frameworks into a holistic mixed-methods evaluation of VAM. By doing so, I aim to highlight issues which prior research has missed or paid insufficient attention to.

I begin by considering VAM from a historical and political perspective through both primary and secondary sources to examine how and why it gained its popularity as well as potential implications. I codify these issues in terms of three core themes—American global power, political strategy and maneuvering, and assimilationism.

Subsequently, I analyze VAM from a quantitative perspective. After discussing quantitative evaluations of VAM in terms of its putative benefits and potential negative ramifications, I turn my attention to the Chetty et al. versus Rothstein debate. These discussions, occurring over a three to four year time period, centered on analyzing one of the key assumptions driving value-added estimates. While it is beyond the scope of this paper to fully discuss the relevant empirical issues, I briefly highlight them to indicate some uncertainty around the statistical or quantitative foundations of value-added estimates.

I then switch to consider Atlanta and Washington DC as case studies of cities where high-stakes value-added systems were actually implemented. By studying these regions, I consider practical

issues that arise with value-added implementations—in the form of a cheating scandal in Atlanta and teacher's union opposition in Washington DC. Finally, I end by considering potential VAM alternatives and establishing a set of policy recommendations on the place of VAM within American teacher evaluation.

2 Context and Historical Background

The Nation at Risk Report under President Ronald Reagan spurred widespread national concern about the state of the United States education system and the extent to which students were underperforming compared to students in other nations. Throughout the late 1980's and early 1990's, public demands for "accountability" increased, beginning with the 1989 Charlottesville Education Summit and the 1994 reauthorization of the Elementary and Secondary Education Act (*ESSA*). These preliminary actions eventually accumulated in the 2001 No Child Left Behind law under George W. Bush—an act that was to have a profound educational impact (Rudalevige, 2003).

While these concerns arose in the legislative sphere, the groundwork for Value-Added Modeling was simultaneously being laid. The Tennessee Department of Education commissioned William Sanders, a statistician at the University of Tennessee, to study student and teacher data beginning in 1982. In particular, Sanders emphasized the effects that teachers could have on students, leading to the development of the Tennessee Value-Added Assessment System (TVAAS) (Aldrich, 2017). Interest in VAM accelerated around 1996 and 1997 when Sanders and others produced papers that demonstrated the high levels of variability in teacher "quality" and the ability to predict future student test scores (National Research Council and National Academy of Education, 2010). While initial pioneers of VAM like Sanders began with fairly cautious recommendations and treated it as a statistical model (i.e. recognizing its underlying assumptions and flaws), these types of

concerns began to dissipate in the 2000's. Articles appeared in venues ranging from *The School Administrator* to *The Los Angeles Times* offering uncritical and unadulterated praise of VAM-based systems (Strauss, 2011). A few years later in 2006-2007, the US Department of Education began allowing districts to move away from the NCLB framework of Adequate Yearly Progress (a model under which a certain percentage of students must attain proficiency on standardized tests) towards growth based models under the Growth Model Pilot Program. Rather than grouping all students together and setting a single standard, these models paid greater attention to what teachers were adding to each individual student's sscores.

More states adopted VAM-based systems under the Obama administration which spearheaded policies such as *ESSA* and *Race to the Top*. *Race to the Top* grants were linked to student assessments, leading many states to begin adopting VAM-based systems in order to evaluate their teachers (Lawson, 2014). Even today, eight states use value added measures as their primary teacher evaluation metric with several others incorporating it into their evaluative frameworks (Klein, 2019). Furthermore, the majority of states not applying VAM use a different metric called student growth percentiles (SGP—a system used by 24 states). SGP models differ from VAM based on technicalities of the type of statistical model that they fit and can sometimes diverge in their evaluations of teachers (particularly in cases of "conditional skewness" which intuitively expresses when teachers have very high or low growth students within their classroom). However, the fundamental idea of assessing teachers based on growth in student test scores remains the same (Kurtz, 2018). Therefore, while I focus on VAM in particular, the analyses in this paper can be extended to other high-stakes standardized testing based models which remain relevant and widespread today.

Due to the proliferation of these models over the past several decades, as well as high-stakes implementations which have tied value-added scores to teacher tenure and salary, it has become imperative to critically assess these evaluative methods. I begin doing so in the next section by

analyzing speeches and written documents during the advent of VAM as well as secondary source analyses of the political factors intertwined with these systems. This political lens provides an essential framework through which to examine VAM which is often treated as value-neutral by many econometricians and policy makers focused on the empirics of these methods. Through conducting this examination, beginning with the 1983 Nation at Risk report during the Cold War, I illustrate that these systems are highly intertwined with certain values and political beliefs as my first primary thesis.

3 Analysis of the Political Rhetoric Around VAM

I construct this analysis through examining the political rhetoric in a few key documents and speeches made in the 1980's through the 2000's. First, I analyze the 1983 Nation at Risk Report. While this report preceded VAM, it incited mass panic over the state of the American education system and eventually spurred the development of accountability-based systems. Subsequently, I analyze the *No Child Left Behind* era which was instrumental in encouraging the accountability movement that further developed during Obama-era policies such as *Race to the Top*. I center my analysis on three core themes: US global power, political maneuvering, and assimilationism. These underlying concepts remain salient throughout the rest of my work as these values remain immanent within VAM, affecting the outcome measures of interest, the real-world implementations, and considerations of potential alternative systems.

3.1 The *A Nation At Risk* Report

3.1.1 The Evaluative Framework of *A Nation at Risk*

T.H. Bell, Ronald Reagan's Secretary of Education, spearheaded the development of the "A Nation at Risk" report through The National Commission on Excellence in Education, citing concerns about "the widespread public perception that something is seriously remiss in our educational system" (The National Commission on Excellence in Education, 1983). The report begins in a drastic and hyperbolic fashion, claiming that "our nation is at risk" and that "the mediocre educational performance that exists today" can be compared to an "act of war" by "an unfriendly foreign power." This pronouncement, grounded in concerns about America's role on the global stage, positions the US in a hypothetical "war" with other nations for educational competitiveness. This can directly be juxtaposed with the real "war" in the form of the Cold War where, in a meaningful sense, the US was engaging in this type of competitive action. However, this framework does not reflect any type of inherent reality as education does not have to operate via a zero-sum game. For instance, an alternative lens could instead construe education as a collaborative venture among nations to best ensure the development and well-being of all students in lieu of positioning children as prospective leaders of a global race for power. Yet, the notion of competing nations on the global stage clearly dominates the National Commission's thinking, for they soon further contextualize that time period as being embedded within the Cold War era, commenting that they have "squandered the gains in student achievement made in the wake of the Sputnik challenge." This mindset directly ties into the development of accountability-driven models, for if the primary goal of education is to ensure that all students possess fundamental skills to guarantee global competitiveness, accountability provides what seems like a viable path for meeting that objective.

Yet, the commission not only emphasizes "a redistribution of trained capability throughout

the globe" and the coming of the "information age" but also the “intellectual, moral, and spiritual strengths of our people.” The report’s authors establish that the global race not only emphasizes economic superiority but also the superiority of core values and beliefs over other nations. This desire to "foster a common culture" regardless of the nation’s "pluralism" or "individual freedom" can again be viewed through the framework of the global Cold War battle against communism and communist systems (The National Commission on Excellence in Education, 1983). While a student’s beliefs or moral values may not be able to be assessed through a standardized test, these notions played a large part in driving the accountability and standardization movements as a backlash to John Dewey’s "Progressive Education." I now pay further attention to the development of these systems through the lens of Andrew Hartman’s research on American education and values in the Cold War which he develops in his 2008 book *Education and the Cold War: The Battle for the American School* (Hartman, 2008).

3.1.2 Cold War Values and Andrew Hartman’s *Education and the Cold War: The Battle for the American School*

Early in his book, Hartman quotes Dwight Eisenhower who, long before the "A Nation at Risk" report, framed education via a militaristic lens as he declared, “No man flying a warplane, no man with a defensive gun in his hand, can possibly be more important than a teacher” (Hartman, 2008, p. 2). While in other eras this quote may not take on any particular significance, Eisenhower’s status as a former general and the recency of World War II imbues it with greater force and strength. As such, long before VAM and "A Nation at Risk," teachers are being portrayed analogously to generals who guide children into a war for global power. This framework easily translates into high-stakes accountability driven systems as there is almost no institution more prototypical of this standardization than the American military. Even the history of standardized testing has its roots

in military service, originating in World War I as a method for assigning jobs to military service members (Fletcher, 2009).

These systems, intertwined with the inculcation and indoctrination of American values in schools, were also based on backlash to an early 1900's oppositional philosophy of education—John Dewey's "progressive" model. In Dewey's philosophy, the core objective of education remains to "dismantle the arbitrary boundaries between the mind and body; between the child and the curriculum; and between the school and society" (Hartman, 2008, p. 12). As such Dewey's framework encourages holistic pedagocical techniques —that education should go beyond academic content to recognize students' autonomy as human beings within the context of a broader society. While there remains controversy over the extent to which Dewey's conception of education actually led to well-rounded individuals rather than "help[ing]" to "prepare obedient and submissive subjects," its philosophical roots were grounded in encouraging student autonomy (Hartman, 2008).

As Hartman comments, Dewey and his ideas turned into a deeply unpopular scapegoat in the Cold War era, for this framework was "deemed too "soft"—politically and epistemologically—for the global struggle against communism" (Hartman, 2008, p. 5). This "counter-progressivism" portrayed students as receptacles for certain values and knowledge that would best enable them to join the global Cold War fight against communism. This notion of education as a measure for "national security," exacerbated by the 1957 Soviet launch of the Sputnik satellite, became a core component of the 1958 National Defense of Education Act (NDEA). A desire for national "security" became one of the key unifying principles in a highly contentious time where according to Senator Listor Hill, American education was wedged between "the Scylla of race and the Charybdis of religion" (Hartman, 2008, p. 185). This analogy—stemming from the Ancient Greek myth of a ship navigating between the six-head monster, Scylla, and the whirlpool, Charybdis—highlights the dire position of American education at that time. Arguments against federal aid came from many

sides ranging from Southern Democrats who opposed distributing aid due to its potential use for desegregation initiatives to Catholic leaders who made claims of discrimination due Catholic schools being ineligible for any of this aid. During this time of strife and disagreement within both the mass public and governmental officials, a common desire for national security became a great unifier, which trickled down to education through a limited and focused law, emphasizing "mathematics, science, and the modern foreign languages."

However, the backlash towards Progressive Education extended beyond the subject content. The types of cultural and value-driven issues at play was synthesized succinctly and vehemently by California Superintendent of Public Instruction Max Rafferty who told Americans

> that our morals are rotten, our world position degenerating so abysmally that a race of lash-driven atheistic peasants can challenge us successfully in our own chosen field of science and our rate of juvenile murder, torture, rape, and perversion so much the highest in the world that it has become an object of shuddering horror to the rest of the human race. (Hartman, 2008, p. 190)

Rafferty not only upheld a racist and xenophobic attitude, grounded in American Exceptionalism and perspectives of Asian nations as consisting of "lash-driven atheistic peasants," but also turned to critique his perceptions of the culture and values of youth during this time. Rafferty portrayed juvenile disrespect and delinquency with the term "slobbism" and sought to instill a "hierarchy of values" enforced by strict discipline in schools, foreshadowing the "zero tolerance" policies of the 1990's (Hartman, 2008, p. 191). These issues again relate directly to the accountability and testing movements as narrow-minded emphases on test scores do enable give teachers to focus on well-rounded or progressive education systems that could encompass concepts such as students' socio-emotional growth. Rather, zero tolerance and high-stakes models of education frequently remove students who are seen as "difficult" from the classroom setting entirely via severe suspension

or expulsion policies, sacrificing those students' education for the sake of conformity and indoctrination. Furthermore, harsh policies are disproportionately applied to students of color and, as such, often drive and exacerbate current educational inequities and achievement gaps (Balingit, 2018). Hartman emphasizes that this agenda was was implemented with "a regime of militaristic individualism," fundamentally irreconcilable with the "social democratic components of progressive education" (Hartman, 2008, p .201). The zero tolerance and accountability movements function symbiotically in order to oppress student autonomy and encourage students to adopt a certain set of values, norms, and behaviors.

Turning back to the "A Nation at Risk" report, these Cold War-based themes, issues, and approaches remain salient throughout the entire report, emphasizing individualism, the desire for a common culture, and the framework of competition with other nations. While significant shifts in education occurred over the 1960's and 1970's—especially over the course of the desegregation movement—many of the fundamental issues remained the same or were still at the forefront of policymakers' minds while crafting education policy.

3.1.3 Evidence and Mandates from *A Nation at Risk*

I conclude the analysis of *A Nation at Risk* by examining the statistics and mandates that the report's authors cite to justify the existence of a problem and drive forward work towards a solution.

In terms of actual findings, the report presents a wide array of statistics on student test scores/outcomes, the expectations of high schools and colleges, the lack of time students spend in school or on schoolwork, and the lack of rigorous teacher qualifications. The authors directly juxtapose these statistics to other nations which perform favorably when compared to the US.

To pursue "excellence" in education, the commission aspires to have "school[s] or college[s] that [set] high expectations and goals for all learners." While this may form their immediate goal

and agenda, they ultimately express loftier ambitions for the creation of "The Learning Society"—in which all individuals dedicate themselves to the lifelong pursuit of learning. While this may at first appear to be a noble goal that maximizes student autonomy, there is a contradiction between this purported goal and the statistics on basic skills that they present, which illustrates the limited conception of what education is and can be from the perspective of the study's authors. Furthermore, they do not allow room for public dissent or disagreement—citing a Gallup poll about how the American public views education as an act of "patriotism" and repeatedly beginning sentences with the word "citizen" to establish education as a way to encourage American values in lieu of an agenda of free, critical thought.

While the commission may rhetorically exert this ethos on the American public, they propose few concrete policy solutions in their report, preferring to cite lofty goals and visions such as "the persistent and authentic American dream" or the "ingenuity of our policymakers,scientists, State and local educators, and scholars" who are tasked with translating these ideas into practice. As such, their grandiose and broadly appealing vision, grounded in concerns about global competitiveness and American values, does not actually establish any key infrastructure or frameworks in order to enact these goals. Furthermore, in line with the Reagan Administration's "New Federalism," the commission was unable to actually implement meaningful education policy at the federal level. Instead, they used their offices as "bully pulpit[s]" to influence state-based actors via fear-driven rhetoric about other nations. These statements especially resonated with Southern states which were undergoing significant economic difficulties throughout the Cold War era. Professors from the University of Virginia and Teachers' College at Columbia traced the shifting goals from the 1970's "from equity to excellence; from needs and access to ability and selectivity; from regulations and enforcement to deregulation; from the common school to parental choice and institutional competition; and from social and welfare concerns to economic and productivity concerns" (Boyd,

1987). With this context, the 1980's can be conceived as a return to ideas that were temporarily set aside during the political battles of the 1960's and 1970's desegregation movement. As such, the "social and welfare concerns" of the 1960's and 1970's as desegregation advocates emphasized equity and students' universal rights to a high-quality education transformed into a narrow-minded focus on "economic and productivity" concerns.

A final development under Reagan was the creation of "statewide performance standards." This infrastructure serves as a precursor for the accountability initiatives in No Child Left Behind (NCLB) and the subsequent adoption of VAM-based systems. Ultimately, the impetus for these systems started with the Cold-War era pushback against Dewey's progressive child-centered education in the early 1900's and a shift away from the equity-oriented desegregation movement in the 1960's and 1970's as concerns about American values and global power gained increasing prevalence. While VAM may, on a surface level, appear more student-centered than universal and unilateral statewide performance standards, it still requires individual students to conform and suppress their individual autonomy and opportunity for free thought. I next turn to see how these notions played out during speeches made in the *NCLB* era.

3.2 No Child Left Behind (*NCLB*)

3.2.1 History, Context, and Values of *NCLB*

The political scientist Andrew Rudalevige tracked the history and political values driving No Child Left Behind as well as its initial results in a 2003 article in *Education Next*. As Rudalevige commented, the key notion underlying NCLB was that of "Accountability" as

> Accountability was hard to be against, but elastic. It served as a way for Democrats to talk about reform without simply talking about increased spending; it was also a sell-

> ing point for additional resources, since ordinarily skeptical Republicans could console themselves that the new funds went to a system newly worthy of investment. While accountability was unproved as a reform tool, there was also no conclusive evidence that it did not work (Rudalevige, 2003)

Rudalevige here highlights that both Democrats and Republicans co-opted this notion of accountability in education, not out of evidence that it would have any tangible benefit for students but rather for their own political purposes. The exact meaning of accountability as well as the split between state and federal control became highly contentious in this debate. During a week—"colorfully called 'hell week'" by a Senate staffer—the entire bill almost collapsed under negotiations about the precise definition of adequate yearly progress. There were tensions with governors who believed the testing requirements to be overly stringent and unfeasible and with civil rights groups that viewed these standards as unjust. Criticisms continued to be lobbied against the bill from all sides—from Conservatives who wanted to include provisions for school vouchers and school choice to teachers unions who opposed the testing requirements.

These tensions ultimately resulted in partially-formed framework in which states were only required to administer the National Assessment for Educational Progress (the *NAEP*) biennially in 4th and 8th grades with no consequences for students' failing and with states being permitted to form their own assessments of student proficiency and standards. As students continually failed to reach state proficiency standards in adequate numbers, states simply lowered those standards with few repercussions despite Secretary of State Robert Paige's labeling of their doing so as "shameful." The significant loopholes and gaps within this legislation indicates that it presented a credit-claiming opportunity for politicians but had few provisions to actually fulfill its intended goals.

These goals and values of No Child Left Behind can be further interrogated through the lens of Paul Shaker and Elizabeth Heilman's critical book *Reclaiming Education for Democracy: Thinking*

Beyond No Child Left Behind. Shaker and Heilman first highlight what can be gleaned about the act even from its title—"No Child Left Behind"—that this phrasing implies "a great competitive race to get a job and get ahead," treating education as a "functional" rather than an "intrinsic good" (Shaker & Heilman, 2008, p. 46). Even if the law was not structurally flawed, the language encapsulates the assimilationism inherent in the act—that it limits the scope of education to providing tangible or material benefits to students. Shaker and Heilman expand on these concerns as NCLB also fails to acknowledge the myriad ways in which children's opportunities are dictated by factors such as poverty, gender, and race, concluding that rather than making "America more equitable and democratic," it may in fact be "more likely to [further] stratify American society" (Shaker & Heilman, 2008, p. 48). This analysis becomes especially essential in light of the educational research which has critically examined the ways in which children's educational experiences differ based on these demographic factors. To name just one example, Stereotype Threat, the notion that students' performance on testing or in the classroom can be altered by stereotypes of how individuals with similar identities should perform, can come into play for students from marginalized communities (The Opportunity Agenda, 2011; Steele & Aronson, 1995). Laws that fail to fully account for students' diverse identities and their differential experiences in both school and home environments is unlikely to adequately serve students' educational needs. While the value-added modeling which arises out of this initiative may statistically individualize students through including demographic factors as covariates in the empirical models, it fails to humanize them in the full complexity of their identities.

Shaker and Heilman further establish the connections between these concepts and the drive for standardized testing through portraying this conception of education "as a scientific...or...credentialing process" that "reduce[s]" challenges such as "improving teaching and learning ...to improving data on student achievement" rather than treating it as a "humanistic endeavor" (Shaker & Heilman,

2008, p. 51). Looking back to the Cold War values and backlash to progressive education systems, this anti-humanistic approach can be conceived as a perpetuation of these earlier systems. Beyond testing and value-added models, this educational ideology seeped into other arenas. For instance, a provision was created inside the *NCLB* to alter the "Federal Education Rights and Privacy Act (FERPA)" in order "to make it easier for public school districts and local law enforcement authorities to share information regarding disciplinary actions and misconduct by students". George Bush attempted to justify these actions in a 1999 speech "entitled 'The true goals of education'" in which he commented that "The days of timid pleading and bargaining and legal haggling with disruptive students must be over" (Shaker & Heilman, 2008, p. 56). Direct parallels can be drawn between this rhetoric and that of the Cold War in which assimilationism and conformity were emphasized in order to combat the rise of "slobbism." Bush refers to students—children who are still in the early stages of their psychological development and who are influenced by a myriad of factors outside of the school environment—analogously to an enemy actor who threatens American national security. Children lack an equal playing field, and to refer to them in terms of "pleading" or "bargaining" or "legal haggling" reflects an adversarial or combative relationship rather than one of support for students' intellectual or psychological development. In another 1999 speech, Bush asserts that rather than teaching "moral puzzles," it is the job of schools to give "moral guidance" and to "encourage clear instruction in right and wrong" (Shaker & Heilman, 2008, p .53). The combination of zero tolerance with an inculcation of certain values indicates a clear drive towards assimilationism—that rather than encouraging critical thought or democratic participation about morality or other issues, policies should be crafted in the pursuit of systemic uniformity and compliance to certain norms. These goals becomes especially disconcerting in light of the disproportionate disciplinary consequences that certain students, such as students of color, still face. Furthermore, there is evidence that these disparities are still highly present within the education system, and, if anything, have

continued to grow over time. (Balingit, 2018).

Viewed through the lens of work from scholars such as Rudalevige, Shaker, and Heilman, No Child Left Behind directly connects to movements towards standardized testing and accountability as well as other policies such as school discipline and zero tolerance. Furthermore, this legislation reflects the overarching values and morals from the Cold War era: global power, assimilationism, and political maneuvering—themes which underlie much of these scholars' analyses. In order to further critically examine the interplay of these values and No Child Left Behind, I analyze a speech and article from two key architects and proponents of the legislation—Republican President George Bush and Democratic Senator Edward Kennedy.

3.3 George Bush's Speech on No Child Left Behind

I first turn to George Bush's 2002 speech on No Child Left Behind (Strauss, 2015). His speech draws from thematic constructs such as global power, competitiveness, and assimilationism as well as reflecting a certain type of self-promoting political rhetoric.

Bush begins his speech by juxtaposing the education of children and the threat of terrorism. Early in his speech, he turns to talk directly to the students in the audience

> And as you know, I've got another challenge, and that's to protect America from evil ones. And I want to assure the seniors and juniors and sophomores here at Hamilton High School that the effort that this great country is engaged in, the effort to defend freedom and to defend our people, the effort to rout out terror wherever it exists, is noble and just and right, and your great country will prevail in this effort. (Applause.)(Strauss, 2015)

In this speech, purportedly about education, Bush turns to vigorously justify his approach to

the war on terror, directly involving the students in the audience as participants in that fight. He employs rhetorical flourish in doing so—emphasizing "noble and just and right" and including the students, both via the direct address and labeling the United States "your" country. There is little room for democratic critique or engagement—if students strive to stand for "freedom," often considered a classic American value, they must too stand with him on his highly controversial approach to handling terrorism, which ultimately led to wars costing trillions of dollars and significant loss of life. In this portion of the speech, Bush places the students in political alignment with himself, both encouraging them to assimilate to certain values and placing educational issues in the midst of a global power dispute.

He later reemphasizes this point, stating that while he "long[s] for peace," he must "lead the world against terror" to ensure that "your children and your grandchildren" will have the same amount of "freedom," which he labels "the precious gift that one generation can pass to the next...and a promise that [he] intend[s] to keep to the American children." A fundamental irony pervades this speech—that Bush encourages freedom while simultaneously implying that students must support Bush's goals and actions in order to be considered good citizens. Bush continues to treat the students like political pawns through his act of making a "promise"—one that may be both politically and rhetorically appealing but which is ultimately impossible to guarantee. In these early sections of Bush's speech, the Cold War and Reagan era values on education continue to percolate in that education is being depicted as intertwined with the quest for global power—a quest which students can partake in through assimilating and conforming to governmental rhetoric and demands.

Bush later pivots from these themes of freedom and global power to discuss accountability and connect it to notions of assimilationism. He comments that "accountability" is the "first principle" as schools must "teach the basics and teach them well" in that "states [must] design accountability

systems to show parents and teachers whether or not children can read and write and add and subtract." While the importance of these basic skills may be almost incontrovertible, Bush rhetorically relies upon this to justify mass standardization. He depicts any in opposition in an overly simplistic light, stating "I understand taking tests aren't fun. Too bad...I need to know whether or not children have got the basic education." As such, Bush fallaciously conflates the acquisition of elementary skills with demands for test-based accountability systems and remains entirely dismissive of the plethora of negative ramifications that test-based systems can have. Furthermore, contradictions abound in Bush's speech as he simultaneously reassures teachers "I trust you" (repeating the phrase three times in a short paragraph) but comments immediately afterwards that they'll spend money on "methods that actually work," implying a lack of trust in teachers and a desire to replace teacher autonomy with universal systems.

Bush's speech ultimately relies on these key notions of global power, assimilation, and political maneuvering. Rhetorically, he repeatedly attempts to emphasize the amount of trust he has in students, teachers, and families and yet simultaneously pushes them towards a conformist and assimilationist framework with a certain set of values. While the emphasis may be on "no child left behind," he fails to acknowledge the ways in which students differ or the structural and systemic barriers that some face (e.g. by race, class, etc.). Bush presents these notions in a simplistic and rhetorically appealing way, which a priori dismisses any potential oppositional arguments.

3.4 Senator Edward Kennedy on No Child Left Behind

However, No Child Left Behind was not solely a Republican-driven bill but rather a piece of bipartisan legislation with advocates within both political parties. One of its key proponents was Democratic Senator Edward Kennedy. Six years after the passage of No Child Left Behind, Kennedy reflected on the bill in a *Washington Post* article (Kennedy, 2008). While he acknowledged some of

the failures of the legislation, he did not express vehement opposition to it or regard it as a failure.

Early in his article, Kennedy comments that one of the core goals of education should be "preparing our citizens to compete and win in the global economy." Similar to other speeches and writings, Kennedy depicts one of the primary goals of education as being about global competitiveness rather than students' individuality or autonomy. While Kennedy does acknowledge that students have different identities—"black or white, immigrant or native-born, rich or poor, disabled or not"— he fails to extrapolate from those labels to the ways in which education systems can systematically underserve certain students. Rather, Kennedy homogenizes these students' identities into a single-minded body with goal of ensuring American success.

A similar dynamic occurs within Kennedy's commentary on test scores which discourage "innovation in the classroom," stating that "a broader array of information beyond test scores" is necessary and that parents and teachers should be given greater room to encourage student success. Yet, Kennedy remains vague about what these additional measures look like and clearly still sees a fundamental role of standardized testing in any educational system. The most concrete policy Kennedy highlights is on school funding as he states that "the law fails to supply the essential resources that schools desperately need to improve their performance." Again, while adequate funding may be necessary for school reform, Kennedy gives it disproportionate focus as an easy solution in lieu of more substantial or controversial alterations that could be made to the education system. He entitles his article "How To Fix 'No Child'," displaying his belief that it can indeed be "fixed" and doesn't need to be entirely replaced or re-envisioned.

Former teacher and education writer Dan Brown critiques Kennedy's article in a reply (Brown, 2008). Brown comments that "schools" have "become corporate-modeled testing factories" that can "make students feel like failures." This critique goes much further than Kennedy's and proposes that it may be time to move towards an entirely different system. Furthermore, Brown states that "some

high schools may tacitly encourage failing students to drop out" in order to not have to count those students' "test scores." As such, these measures purportedly intended to ensure that all children are able to achieve some measure of "success" may ultimately undermine that goal in practice. Finally, Brown puts forth an array of proposals from "small class sizes" to "opportunities for small-group support" to "mental health services" to "a diverse curriculum." These policies all acknowledge students as individuals who have experiences outside of the school environment that may impede their academic performance and which I return to later when reflecting on possible alternatives to VAM and high-stakes testing environments.

3.5 Conclusion—The Political Consequences of VAM

The Obama Administration's *Race to the Top* program that dispensed funding based upon students' academic performance only served to further encourage states to adopt accountability systems like VAM for teacher evaluation (Lawson, 2014). Furthermore, many states still employ value-added models or similar systems based on analyzing the impact of teachers on student performance. As such, questions surrounding the effectiveness of VAM and possible alternatives continue to remain salient within the education policy landscape.

In summary, value-added modeling was grounded in a system of political negotiations that involved numerous compromises from legislators with differing agendas. Furthermore, these systems cannot be extricated from the Cold War era values and morals about global competitiveness and assimilating students into American society which pervade much of the rhetoric used to describe them. While many studies or examinations of VAM treat it as a purely quantitative or objective model, it is essential to acknowledge and recognize political and historical roots which were grounded in a certain vision of what education ought to look like. It is indubitable that if education in the US were based on other models such as Dewey's progressive vision, different and more holistic systems

would have evolved for assessing students and teachers.

In the next section, I turn to examine value-added modeling from an entirely different, quantitative perspective. In the first part, I ask from an evidence-driven perspective, what are the putative benefits of systems like value-added modeling (many of which are tied to students' future academic success or earning potential)? Subsequently, I discuss evidence on the potential drawbacks and limitations of standardized testing based systems in terms of potential consequences on student well-being and growth. While evaluating these components of value-added modeling, I keep in mind the underlying values and morals which I have uncovered as well as VAM's historical context. In the second part of the next section, I question whether, under the temporary assumption that these test scores are in fact a good or useful metric, they can actually be estimated accurately. In order to do so, I briefly look at the discussion between scholars Raj Chetty and Jesse Rothstein on one of the key statistical assumptions underlying value-added models.

4 Quantitative Evaluations of VAM

In this section, I move from examining the political ramifications of VAM through the lenses of US global power, political maneuvering, and assimilationism to examining the potential benfits for VAM as well as one of the key quantitative issues behind value-added estimation: students' possible non-random assignment to teachers. I engage in this subject primarily through the lens of two of the most prominent economists conducting work on VAM—Raj Chetty and Jesse Rothstein— who have debated on this topic by invoking quasi-experimental estimation techniques.

I begin this section by discussing this topic at a general and less technical level. Firstly, I look at some of the evidence that econometricians have put forth in favor of high-VAM teachers in terms of putative benefits to students (in terms of salary, future academic outcomes, etc.). However, I

tie these benefits back into the ideas of assimilationism and a limited perspective of what student achievement looks like discussed in Section 3 as well as whether alternative methods could be expected to achieve similar results. I do not explore the technical methodologies used in order to estimate these future outcomes for students within the scope of this paper but rather consider the implications if they are taken on face value (however, there is room for future work to also delve deeper into assessing the estimates of projected student outcomes on a more technical level).

Next, I provide a general, non-technical overview of the main issues discussed by Chetty and Rothstein of student allocation to teachers. In this section, I also consider some of the literature on how students are actually assigned to teachers in order to frame Chetty and Rothstein's argument. I stress that Chetty and Rothstein's discussion is on just *one* of the technical problems with VAM, and there are several others which I do not examine in detail within the scope of this paper. However, in my final section, when I discuss directions for future research, I will return to very briefly highlight some of these issues.

4.1 General Overview of Benefits and Limitations of VAM

I begin by providing a high-level overview of VAM from a non-technical perspective—starting with the putative benefits and then followed by the limitations as discussed by Chetty and Rothstein.

4.1.1 Benefits of VAM

One of the most prominent studies in favor of VAM was conducted by Chetty, Friedman, and Rockoff (2014b) in which they implement two strategies for estimating long-term impacts of VAM—one based on controlling for student demographics and other characteristics and another based on a quasi-experimental design strategy. The key premise of this second strategy was to look at high value-added teachers who leave schools and are replaced with lower VA teachers. Thus, this change

in teacher quality (as measured by value-added scores) can then be cross-compared with long-term student outcomes in order to generate causal estimates of the effects of having high VAM score teachers on students' future trajectories.

Based on both of these strategies, Chetty, Friedman, and Rockoff identify a number of key benefits of VAM in terms of long term outcomes. These include, based on a one standard deviation increase in a teacher's VAM score for a single grade: a 0.82-0.86 percentage point increase in the likelihood of attending college at age 20 (and attendance at "higher quality" colleges based on graduates' earnings), 1.3% higher annual earnings on average (translating to approximately $39,000 over the course of the students' working life), reductions in teenage pregnancy, increases in neighborhood "quality" (measured by the percentage of college graduates in that neighborhood), and increases in participation in 401(k) retirement savings plans. Furthermore, they extrapolate to the classroom level and find increases in terms of classroom lifetime earnings for a given teacher of $185,000-$250,000 on average. (Chetty et al., 2014b, p. 2634-2636).

4.1.2 Problematizing the Benefits of VAM

Even if Chetty, Friedman, and Rockoff's estimates hold validity, there are several levers through which I can critique or problematize the implications of those projected outcomes.

First, I question the benefits studied in terms of the political framework and analysis I laid out in the previous section. While undoubtedly, metrics such as income, savings, and further education do have importance in terms of students' quality of life, they put forth a limited view of what education ought to do. There may be numerous other metrics or conceptualizations of education from skills such as grit, conscientiousness, or critical thinking to evaluations of student well-being that are neglected in this model. As such, the metrics applied by Chetty, Friedman, and Rockoff may not actually be the most relevant in terms of reflecting a wider sense of student well-being

either on either individual or societal levels (i.e. in terms of having citizens who can actively reflect on and engage in political and social processes to better society). In fact, these measures trend towards pushing students towards a version of assimilationism through implying the aspirations students ought to hold.

Furthermore, recently, education scholars have begun to pay greater attention to so-called "soft" or non-cognitive skills that may also be meaningful for students. This research gives way to a few fundamental questions—Can VAM estimates help predict any of these soft skill outcomes? Are teachers able to actively impact any of these non-cognitive impacts? And, finally, is there any sense in which these non-cognitive skills are meaningful to students and their well-being beyond the effects predicted by VAM?

Research on these questions so far has been limited. However, Kirabo Jackson in a recent 2019 article examined the correlations between teachers who improve student behavior (measured by metrics such as suspensions, GPA, and progression to the next grade) and teachers' value-added scores (K. Jackson, 2019). While behavior certainly does not capture all non-cognitive skills, it provides an indicator of some skills beyond standardized test scores and is far more easily calculated from available data than more abstract notions such as student independence or conscientiousness.

In addressing this question, Jackson found that while there was some degree of correlation between what he termed the "behavior index" (the composition of all of student behavior metrics) and value-added scores, "effectiveness along one dimension is a poor predictor of the other." In quantifying this correlation, Jackson found that when he created a separate value-added score for student behavior, only around 40% of the teachers in the bottom third of behavior value-added scores had above average test value-added scores. Analogously, for the top third of behavior value-added scores, only around 58% of those teachers had above average test-based value-added scores. From this data, there is a clear disconnect when evaluating teachers based on tests as compared to

other metrics, and one is not necessarily clearly predictive of the other.

Sarah Fleche (2017) also examines the issue of non-cognitive behaviors within the context of schools in the UK. Compared to Jackson, her dataset consists of more finely tuned data on students. She splits non-cognitive skills for students into internalizing behaviors (those that are inward facing such as students' level of worry/nervousness, how solitary they are, signs of nervousness, etc) and externalizing behaviors (such as hyperactivity, restlessness, or bullying). This data is reported by students and parents, and similarly to Jackson's study, she finds low levels of correlation between value-added test scores and behavioral outcomes (correlations of 0.08 to 0.14 between math value-added scores and non-cognitive behaviors). By contrast, teachers' effects on internalizing behaviors had a statistically significant correlation on externalizing behaviors of 0.54, indicating that the behavioral indices are highly related; whereas, there is not a significant connection between behavioral and academic outcomes (Flèche, 2017).

Both of these studies imply that there are certain non-cognitive skills that VAM is unable to clearly capture. A second question is whether teachers are able to actively influence these skills (as, if not, there would be little reason to assess teachers based on them). Both of the aforementioned studies also investigate this issue. Jackson describes the magnitude of the effects teachers can have on behavioral measures by comparing a teacher at the 85th percentile to one at the 50th percentile along these metrics. For a math teacher, he finds that this comparison means a 1.2% decrease in the likelihood of being suspended, no effect on absences, a 0.063 GPA increase, and a 2.64 percentage point increase in on-time grade progression (C. K. Jackson, 2018, p .2090). The effects from English teachers were similar although smaller in magnitude. Fleche had similar results in her paper. While it is less immediately interpretable, she compares teachers in the 75th percentile to those in the 25th percentile, finding that it corresponds to an increase of 1 point on a 0-20 scale for internalizing behaviors and 0.5 points on a 0-20 scale for externalizing behaviors. Furthermore,

these are comparisons across a span of four years (from UK year 3 to 6 which is equivalent to comparing US grade 2 to grade 5). As such, both studies indicate that student behaviors are malleable across contexts (Jackson's study was across 9th graders in North Carolina while Fleche's was for primary/elementary school students in the UK).

Finally, I turn to the third question of the potential impacts of these non-cognitive skills on students. Both Jackson and Fleche consider these in terms of the same types of tangible outcomes that Chetty, Friedman, and Rockoff discussed. Jackson finds that, in fact, teachers' value-added scores on student behavior metrics were significantly more predictive of all the tangible outcomes he examined when compared to value-added test scores. As one example, teachers in the 85th percentile of behavior value-added scores improved their students' chances of on time high school graduation by 1.46% percentage points relative to the average teacher for behavior value-added scores. By contrast, teachers in the 85th percentile of test value-added scores only added 0.12 percentage points to their students' chances of on-time graduation when compared to average students on this metric. Another way of viewing this same data is that among the 10% of teachers who have the biggest impact on high school graduation rates, 93% also were in the top 10% for behavior-based value-added scores while only 20% were in the top 10% of test-based value-added scores. These discrepancies appear in which behavior value-added scores are approximately 10 times more predictive of student outcomes follow for several other outcomes as well including increases from 9th grade to 12th grade GPA, the probability of taking the SAT, and the probability of intending to attend college.

While Fleche did not find that behavioral value-added effects were significantly greater in magnitude than those based on test scores, she did find that behavior-based effects were significant and independent of test-based effects. The first effect she examined was higher education attendance in which she found that a one standard deviation increase in value-added scores for math increased the chance of higher education attendance by 0.98 percentage points. For internalizing value-added

scores, a one standard deviation increase resulted in an increase of 0.55 percentage points while a one standard deviation increase for externalizing value-added scores resulted in an increase of 0.86 percentage points in the probability of higher education attendance. Including both effects at the same time in the model, while controlling for key demographic variables, illustrates that these effects operate independently of one another and do not cancel each other out. Fleche then explores earnings at age 20 and finds similar results with one standard deviation value-added increases leading to salary increases of £480 for math value added scores, £132 for internalizing value-added scores, and £487 for externalizing value-added scores (equivalent to approximately $626, $172, and $635 respectively). Once again, when both effects are inserted into a single model, they both remain significant. Furthermore, put in other terms, a teacher in the bottom 10% in terms of externalizing value-added scores is equivalent to a student missing a third of a school year in terms of impacts on future earnings.

Finally, Fleche also addresses the phenomenon known as "fade-out" in terms of both test-based value-added scores (particularly for math) and in terms of non-cognitive value-added scores. For effects on math scores, Fleche finds that they fade out over time—that teachers' positive effects on students' math scores decrease over time and then the positive effects reemerge in adulthood. Conversely, non-cognitive skills are persistent meaning that students maintain those non-cognitive skills in future years. This provides another argument for including non-cognitive skills as they don't suffer nearly as much from these fade-out effects.

These studies center both on concrete skills/behaviors as well as tangible benefits and outcomes (earnings, education performance, and higher education attendance) of the type that Chetty, Friedman, and Rockoff also evaluated. However, once again, as I have pointed out, these effects pre-suppose a very limited view of what education should be and of what adult success ought to look like. There are numerous other so-called soft skills which may be valuable to teach in educational

contexts and which may have both tangible and intangible benefits to students. As one of many possible examples, I select emotional intelligence as a skill which may be less easily assessed but which may be beneficial to students.

Pablo Fernández-Berrocal and Desiree Ruiz (2008) provided a literature review of emotional intelligence in education in terms of the extent to which it can be taught and its potential benefits to students. They summarize four key benefits to increased emotional intelligence for students: 1) Improved interpersonal relationships, 2) Higher levels of psychological well-being (as measured by factors such as social anxiety, depression, and reactions to stress), 3) Increased academic performance (as higher emotional intelligence enables students to better engage with academic tasks), and 4) Lower levels of disruptive or antisocial behavior (such as substance abuse). Furthermore, studies have been conducted which have found that successful programs have been implemented in educational settings for improving students' levels of emotional intelligence.

Again, emotional intelligence is just one of many examples of so-called soft-skills that present both tangible and intangible benefits to students. Some of these (such as improvements to interpersonal relationships, rates of substance abuse, or response to stress) have not traditionally been under the purview of metrics that econometricians and education policy analysts have used to assess the success of teachers or schools. Yet, these factors are undoubtedly essential to student well-being and to students flourishing as adults and contributing successfully to society. Therefore, it is worthwhile to further question why the types of attributes assessed by Chetty, Friedman, and Rockoff are the ones tracked in order to evaluate teachers and the educational system more broadly. One reason is undoubtedly the simple fact that the data is far easier to track and more available than other types of data (i.e. interpersonal skills can be more challenging to assess and analyze on a large scale). Yet, beyond data availability, it also reflects a certain set of values placed on the education system in terms of students being able to assimilate into mainstream society.

Ultimately, while there is very strong evidence stating that having a high value-added teacher does have some causal effects on students' future outcomes, more recent data has illustrated that there may be other non-cognitive outcomes that have just as strong (and potentially much stronger) effects which are independent of those estimated via value-added modeling. Chamberlain actually quantified the impact that teachers have on college attendance that can be attributed to test scores as opposed to other factors for teachers (2013). When constructing a factor for teachers in his analysis based solely on test score data, he found an increase in one standard deviation in that factor led to a 0.13 percentage point increase in the chance of college attendance. However, when it was constructed via college attendance data instead, he found a one standard deviation increase on the teacher factor to lead to a 0.79 percentage point increase in the probability of attending college. In other words, from this data, only around 16% of the total impact of teachers on students' future college attendance can be attributed to test scores alone with the rest being attributable to other factors. With these results, test scores are likely overweighted in current models of teacher evaluation while other factors have traditionally been given relatively little attention and weight despite the significance of their impacts. Even though the entire teacher factor may appear relatively small in terms of substantive significant (0.79 percentage points being a fairly small increase), it is more meaningful when generalized over many thousand students.

While our main concern in problematizing the benefits of VAM discussed by Chetty, Friedman, and Rockoff (2014b) was to highlight the limited view of education their models endorse and the numerous missing factors that can also influence students and to which too little attention is paid, there are two other issues with accepting the benefits that they highlight on face value. The first, which I already briefly discussed, is the problem of so-called "fade-out" which Chetty, Friedman, and Rockoff also explain in their paper. They find that benefits in students' scores diminish over a period of 3-4 years and ultimately settle at a level of around 25% of the initial impact. Despite these

benefits diminishing, they arise again in adulthood where students are able to receive all the benefits they analyze (such as increased income, better savings, etc). However, if other non-cognitive skills don't have this fade-out effect, that is one reason to potentially prefer them as a metric over test scores.

Secondly, there are questions about how these benefits should be linked to actual policy. For instance, Chetty, Friedman, and Rockoff find some evidence of cheating/score manipulation in the top 1% of VAM scores (patterns of high rises in test scores followed by low ones). I don't delve too deeply into these issues here other than to highlight the quantitative evidence for them as I explore problems with high-stakes evaluative systems further in Section 5 where I consider both Washington DC and Atlanta Public Schools as case studies of what can go poorly in practice.

From this analysis, I do not necessarily conclude that VAM is an entirely meaningless metric or one that should be dismissed entirely (although discussing the technical merits and deficiencies of the estimates of VAM benefits is something that should also be analyzed further in future research). Rather, I urge caution in how much weight it is given in teacher evaluation formulas and in how it is actually implemented (discussed further in Section 5). Furthermore, education has the potential to provide more to students than solely tangible outcomes in terms of dividends to income, college attendance, etc, and it is not at all apparent that a system based around standardized testing will help students with these soft skills or elements of personal well-being in the way that other metrics might. A bolder vision can be constructed for an education system that empowers students to do more than to simply assimilate into society as it currently is, enabling them to partake in critical thinking, meaningful relationship building, civic engagement, etc.

4.1.3 A High Level Overview of the Technical Issues within VAM

In this section, I provide a high-level overview of the Chetty et al versus Rothstein debate in terms of its fundamental premise (the effects of how students are allocated to teachers to VAM estimates) and scholarly work that delves into that issue.

In a 2016 article for *IRLE* (The Institute for Research on Labor and Employment), Rothstein highlights the key premises of Chetty et al's research and the major issues that he identifies with it (Rothstein, 2016). In that article, Rothstein states that there may be biased estimates in VAM scores as a result of non-random assignment of students to teachers. Rothstein surmises that these "assignments are complex processes, incorporating parental requests, teacher specializations, and assessments of students' individual needs and social dynamics" (2016). Rothstein indicates that this fact presents a problem for many studies of VAM, and that he finds quantitative evidence for this bias in value-added scores. He conducts this work by analyzing the effect sizes of 5th grade teachers on 3rd and 4th grade test scores after controlling for the covariates present in the value-added model. Rothstein found significant "effects," which are clearly invalid (as 5th grade teachers cannot retroactively alter students' 3rd and 4th grade test scores). As such, value-added modeling is not capturing all facets of student assignment into classrooms, resulting in biased estimates.

Other research has also examined the question of student assignment to classrooms. For instance, Dieterle, Guarino, Reckase, and Wooldridge also examined this same question (2013). They find that there are significant amounts of bias in assignment of students to classrooms, especially based on student test scores. Furthermore, these discrepancies vary greatly by school and make value-added score estimation especially challenging.

This problem remained mostly unexamined until Chetty, Friedman, and Rockoff conducted an innovative study in order to estimate the magnitude of these biases (which prior studies had been unable to successfully quantify) (2014a). This study relied on a quasi-experimental design based on

examining changes in average value-added scores when teachers enter or exit a school. From over 10 years of data in New York City schools with thousands of changes in teachers, their study finds little evidence of biased VAM scores (the estimated bias is sufficiently low that it can essentially be ignored).

However, Rothstein problematizes Chetty et al's empirical strategy as it relies on the assumption of teacher switches also being random and disconnected from any other changes in student performance by school and grade. Rothstein's research tests this claim and finds it to fail in practice. Namely, teacher switching tends to vary based on the previous teacher's value-added scores and the scores of the incoming cohort of students. As Rothstein indicates, when students have higher test scores than the previous cohort, high-value added teachers tend to replace lower value-added teachers. Conversely, when students have lower test scores than the previous cohort, low value-added teachers tend to replace higher value-added teachers. If true, this would pose a major issue for Chetty, Friedman, and Rockoff's strategy for estimating the bias present in value-added scores.

While Chetty, Friedman, and Rockoff responded to Rothstein that Rothstein's results are merely based on "mechanical effects"—in that they simply relate to how teachers' value-added scores are calculated— Rothstein disputes that claim in two ways. First, Rothstein employed strategies to eliminate these sources of bias, and, second, he also found statistically significant correlations between teacher switching and certain student attributes like race and free lunch status that weren't used for value-added estimation. If Chetty, Friedman, and Rockoff's response was in fact valid, Rothstein argues that none of these correlations would stay present. Furthermore, when Rothstein employs strategies to attempt to estimate bias in value-added scores by adjusting for the relationship between students' prior scores and incoming teacher value-added scores, he finds a much higher level of bias than that found by Chetty, Friedman, and Rockoff (up to one third of the

variability in teachers' value-added scores).

With that magnitude of bias, it is no longer ignorable and can lead to serious errors in assigning value-added scores to teachers. Furthermore, it also calls into question Chetty, Friedman, and Rockoff's second paper on the effects of high value-added teachers on outcomes like college attendance and earnings that I discussed in the prior subsection (2014b). When Rothstein attempted to replicate the results of that paper after taking into account the bias due to non-random assignment of teachers, he found that these estimates of the value-added benefits were also greatly reduced, and many were no longer statistically significant.

While it is beyond the scope of this paper to fully analyze the empirical strategies employed by Chetty, Friedman, and Rockoff or by Rothstein, and it is worth noting that there was significant debate on both sides of this issue, Rothstein's critique at minimum illustrates that Chetty, Friedman, and Rockoff's results are not indisputable, and serious concerns remain about the extent of bias in value-added scores due to non-random assignment of students to teachers. There is room for future research to apply new data and empirical strategies to this problem to further attempt to quantify its magnitude.

4.2 Conclusion of the Quantitative Analysis

In this section, I have identified issues with the quantitative benefits of value-added models as well as critiques of research claiming that any bias in these models can be treated as ignorable. In terms of the benefits, it is not clear that they are as high in magnitude as Chetty, Friedman, and Rockoff claim; furthermore, there is preliminary research on the promising effects teachers can exert on students via other characteristics such as student behavior. Value-added modeling also fails to capture many so-called "soft skills," which are also important for students' well-being and development.

In terms of the quantitative issues, Rothstein's research highlights a serious critique which implies that there may be significant inaccuracies in estimates of value-added scores. It is also worth emphasizing that bias from student assignment to teachers is not the sole source of bias or quantitative issue in the field of value-added estimation. Other possible issues that I did not discuss in this paper range from measurement error on standardized tests to different functional forms/types of value-added models (as the statistical techniques for value-added modeling frequently vary by state and district) to the notion of peer effects (the idea that students are treated as independent when they may exert effects on one another). Many of these are still areas of active research in terms of estimating the extent to which they may also affect value-added scores. As such, even if I ignore all of the qualitative problems with value-added scores, it is not entirely clear how good they are as a quantitative metric.

5 Case Studies of VAM Implementations

In this section, I transition from viewing VAM from a quantitative perspective in order to examine real-world implementations of VAM or similar systems. Regardless of the political implications of VAM examined in Section 3 or the quantitative validity in Section 4, it is instructive to analyze actual instances of VAM as frequently even policy proposals that seem theoretically beneficial encounter real-world resistance or other challenges that prevent those benefits from being fully realized. Hudson, Hunter, and Peckham (2019) explore this issue in a recent article and explain some reasons why policy implementations may fail to achieve their full potential. This framework provides a useful framework for the case study regions of Washington DC and Atlanta.

In their article, Hudson, Hunter, and Peckham identify four major threats to a given policy's success. The first is overoptimism in terms of expectations for the policy which can be unraveled

due to factors such as policymakers misunderstanding stakeholders, lack of evidence, or challenges in delivering the goals of the policy. The second is dispersed governance where policies that are successful in one location may not necessarily have the same benefits in another. Furthermore, there may be a disconnect between local actors and those creating the policy at a higher level, resulting in confusion and alterations to the policy as originally intended. Third, there may be insufficient collaboration among relevant stakeholders. For a policy to succeed, it does not necessarily require unanimous agreement (an unachievable goal), but it does demand a constructive handling of differences and negotation throughout both "vertical and horizontal" organizational structures. Finally, there are problems with the so-called "political cycle" in which politicians desire to implement policy for the short term in order to claim credit but are less interested in handling all of the long-term intricacies which are necessary for the policy to succeed. Hudson, Hunter, and Peckham suggest policy support programs in order to tackle these issues in which greater care is given both to the initial creation of the policy and to subsequent review of which policy milestones are met by particular deadlines and how exactly the implementation functions in practice. Several of these issues become apparent in the Washington DC and Atlanta which I examine; thus, it provides a useful overall framing as I seek to evaluate and assess these case study regions.

In order to examine this, I turn to two case studies: Washington DC under Michelle Rhee (resulting in opposition by Teachers Unions) and the Atlanta Public Schools cheating scandal. While both cases may be somewhat extreme and may not necessarily generalize throughout the entire US, they are illustrative of some of the problems that systems like VAM might face—in the first case, via opposition movements (by teachers, parents, or other community members) and, in the second, by attempted subversion. These problems have frequently arisen in the context of other education policy ideas and contexts (busing provides one prominent example). Furthermore, beyond the problems arising in Washington DC and Atlanta, there are other variable factors and challenges

with VAM such as the need to evaluate and assess students in subjects and grade levels for which it is impossible to create standardized tests. I briefly address some of these broader issues at the conclusion of this section.

5.1 Case Study 1: Washington DC Under Michelle Rhee

I begin by analyzing the first case study—Washington DC under Michelle Rhee and the high-stakes system that resulted (known as IMPACT). I begin by describing the IMPACT system and the tensions between Rhee's government and teacher's unions (two of the most powerful stakeholders in education in Washington DC at the time). I then turn to discuss some of the benefits of the IMPACT system and the broader lessons that can be gleaned from this case study before moving onto the second case study of Atlanta Public Schools and the subsequent cheating scandal.

5.1.1 Description of IMPACT and Tensions with Teachers Unions

Laura Winig (2012) produced a detailed case study summarizing the key issues with the DC education before and during Rhee's tenure. As Winig described, Rhee inherited a system in deep disarray which according to standardized test scores from the 1950's and 1960's. By 1995, the education system had reached a "state of emergency," and by 2007 Washington D.C. was last in terms of math scores, second to last in reading, and had a high school graduation rate of 43% in spite of a billion dollar budget (equating to the third highest level of spending on a per pupil basis in the US).

As such, the newly appointed mayor Adrian Fenty designated Rhee as the school chancellor in June 2007 with the goal of making drastic changes to DC's education system. Among the many overhauls she made (closing schools, firing teachers and principals, etc) was the design of a teacher evaluation system known as IMPACT.

This system consisted of several core components—the first (comprising 50% of total IMPACT

scores) was a value-added model based on the Washington D.C. standardized tests (the D.C. Comprehensive Assessment System (DC-CAS)). This only applied to the 20% of teachers in the system who taught in grades or classes that implemented these standardized tests. The second component was known as the Teaching and Learning Framework (35% of the IMPACT score) which consisted of five sets of classroom observations—three by an administrator (principal or assistant principal) and twice by an outside "master evaluator." The third component (10 % of the IMPACT score) was Commitment to the School Community which assessed teachers' involvement beyond the classroom. The remaining 5% was on school-wide value-added data (once again based on the DC-CAS standardized testing). For the 80% of teachers who did not teach classes with end of grade DC-CAS tests, the weighting was redone so that tests other than DC-CAS made up only 10% of the teacher's total IMPACT score with the vast majority of the score (75%) stemming from the observations conducted via the Teaching and Learning Framework.

This system, which enabled the district to fire teachers rated minimally effective immediately and which hired the "master evaluators" with no union input, was immediately and vigorously opposed by the Washington Teacher's Union (WTU). Rhee refused to make any concessions such as piloting or rolling out the system gradually, and it took over a year for the union vote to agree to the new system which was mainly a result of high bonuses for highly effective teachers (up to $25,000 with large base salary increases for teachers who were repeatedly assessed as highly effective). Minimally effective teachers were given a one year time span to improve while ineffective teachers were immediately fired. Despite the significant initial margin of approval for the IMPACT system (1412 votes to 425), support diminished when it was announced that 5% of the teachers in the system would be fired and another 17% were on a one-year warning to improve their IMPACT scores.

Winig (2012) cites blog posts of teacher reactions in the wake of IMPACT from teachers who felt

the changes to be overly rapid and punitive. Among the concerns listed were pressure for teachers to cheat the system, lack of consideration for students' sociodemographic factors over which teachers exerted little control, an emphasis on "teaching to the test," and the effects of having more students or more disruptive students in a given classroom on the overall VAM scores. While the IMPACT system continued after Fenty lost re-election in 2010, Jason Kamras (who had helped Rhee design the initial systems) solicited significant teacher feedback via focus groups in order to improve it substantially.

IMPACT has continued to this day and is still generating controversy and questions surrounding its future in DC and elsewhere. A recent *Washington Post* article discusses this controversy and that the current DC School Chancellor Lewis D. Ferebee is launching a widespread, systemic evaluation of the IMPACT system (Stein, 2019). Not only did tensions grow between teachers and the school system, but Washington DC also experienced a scandal in 2018 in which schools graduated students who did not meet district requirements (1 in 3 graduates did not meet these requirements such as too many class absences or improperly taking make-up classes). An extreme case of this occurred at Ballou High School in one of DC's poorest neighborhoods where all the graduates were accepted into college for the first time, yet it was subsequently found that 3 out of 5 students did not meet the requirements to graduate (Mattingly, 2018). While there have been some improvements to the issues present in IMPACT such as a recent increase of 8% in the teacher retention rate, there has also been significant pressure on teachers to simply pass students in order to meet requirements.

Mattingly (2018) further discusses this in his article on IMPACT and the fact that Jason Kamras, while one of the pioneers of IMPACT, refuses to implement it in the Richmond Public School system (where he currently serves as superintendent). Mattingly describes teachers and parents who have indicated that the IMPACT system has created a "culture of fear" and that

teacher retention is still very low within the lowest performing schools (1 in 3 teachers leaving each year). In a survey, half of all teachers felt pressured to pass students regardless of their performance, and some cited the fact that they felt forced to no longer operate in the students' best interests due to the evaluation system. Beyond, Mattingly in Richmond, Virginia as a whole has recently been considering proposals to enable school districts to reduce the 40% weight currently given to standardized tests and increase the weight on other metrics for teachers such as professional knowledge, professionalism, and instructional delivery (Mercury & Hankerson, 2019).

These issues in DC can be tied back to the broader policy evaluation framework proposed by Hudson, Hunter, and Peckham. The DC IMPACT system meets several of their criteria for dysfunctional policies. For instance, Rhee exhibited over-optimism in portraying teachers as entirely in support of IMPACT when there was actually significant controversy over the policies in which policymakers failed to foresee the ways in which schools might try to evade the system (i.e. by feeling pressured to pass and graduate their students). Furthermore, there was insufficient collaboration in which the teacher's union wasn not included in any of the negotiations and, despite eventually conceding to the policy, disliked it when its actual consequences (i.e. firings and warnings) came to be realized. Finally, political cycle problems were present in which Rhee emphasized immediate gains and moving as quickly as possible to implement drastic, wide-scale changes to the system over creating a policy which would have long-term success. There was little in the way of the policy supports that Hudson, Hunter, and Peckham propose in terms of a longer term vision for policy development and implementation and continual feedback cycles/verifying that the policy was meeting its milestones (until Kamras began a more extensive feedback process after Rhee's departure in 2010). This led to a policy which did have many negative repercussions and which now over 10 years later is due for significant overhaul.

5.1.2 Benefits of IMPACT and Lessons Learned from DC

Despite the negative ramifications of the IMPACT system, there have been some studies which have also highlighted its benefits. Since its initial implementation, it has been the subject of study from scholars from Stanford, Brown, and the University of Virginia. Audrey Breen (2019) summarizes these studies which declare IMPACT to be highly successful.

As Breen summarizes, two early studies of IMPACT found the system to encourage ineffective teachers to either leave the system or improve their teaching skills. Furthermore, ineffective teachers (according to their IMPACT scores) who left were replaced with ones who were evaluated as more effective at increasing student achievement. More recent studies of IMPACT have essentially come to the same conclusion, finding that only 10% of highly effective teachers left the system in 2013-2017 as compared to 55% of Minimally Effective teachers.

How is it possible to reconcile the tensions with teachers' unions and the pressure to pass and graduate failing students with the high levels of praise from education policy researchers? Once more, this can be seen as a gap between policy theory and practice where IMPACT does produce some positive results through a heavy weighting on test scores yet may not ultimately be worth those dividends due to the culture of fear and teachers who feel as if they no longer can represent their students' interests. Furthermore, while I have not found evidence as such, it would be unsurprising if some of the classroom observations coming from academic administrators could also be inflated in order to improve the appearance of their school's performance. Mark Dynarski (2016) from Brookings Institution has also described mixed evidence on the effectiveness of classroom observations as a metric for teacher performance. Dynarski summarizes research that finds that teacher observations are neither correlated with test scores nor with non-cognitive skills such as "grit" or a "growth mindset." Thus, there are questions about the basis used for teacher observations and whether they they are applied in an effective way by either school administrators

or "master evaluators" (furthermore, they are highly time consuming and financially costly which means that it is essential to ensure their effectiveness).

Ultimately, while I do not unilaterally critique IMPACT, it does possess some fundamental problems in terms of both test scores and observation components. It may be possible to take some of the benefits of IMPACT without the culture of fear and strong opposition from teachers through an alternative system. Furthermore, it may be possible to create similar systems but to reduce the emphasis on test scores as compared to the status quo. This is a question that I further explore in Section 6 where I look at potential evaluative systems employed both in other nations and proposed by education policy experts within the US.

5.2 Case Study 2: Score Manipulation and Atlanta Public Schools

Now, I turn to the second case study of a high-stakes accountability system—Atlanta Public Schools which was subject to a widespread cheating scandal in an attempt to inflate test scores and circumvent accountability measures. While these issues also arise within the Washington D.C. context, they are even more widespread in Atlanta. Yet, there are clear comparisons to be made between the two, and dynamics at play in Atlanta also fit into the policy failure framework put forth by Hudson, Hunter, and Peckham (2019).

Robert Simons and Natalie Kindred (2017) wrote a case study summarizing the issues present in Atlanta Public Schools that led to data manipulation and cheating. Like DC, when Dr. Beverly Hall took over as superintendent in 1999, Atlanta was suffering from significant problems in its public education system such as massive disparities in outcomes by race and income, low SAT scores, a 14% failure of basic proficiency tests required of high school seniors to graduate, and eighth grade math and reading scores at the 43rd and 35th percentile respectively when compared to national scores.

There was national pressure for schools to meet their Adequate Yearly Progress (AYP) targets, and, as such, Dr. Hall set strict target goals within Atlanta and took a data-driven approach to assessing education. She both gave bonuses to all employees at schools which were meeting at least 70% of their district targets while implementing severe repercussions for schools that failed to do so. Over the course of her leadership. Dr. Hall replaced 90% of school principals who were failing to meet targets, and student test scores comprised 25% of principals' evaluations.

This system led to a "culture of fear and a conspiracy of silence" as schools began to achieve unrealistically high gains in student achievement (Simons & Kindred, 2017, p. 19). An elaborate operation to cover up evidence of any wrongdoing occurred for several years until Hall was eventually forced to concede to order an independent investigation in 2009. Even then it ultimately took a statewide investigation, including by the Governor, in order to fully document the extent of the cheating and attribute it to three primary causes—including "unreasonable pressure on educators," a "culture of fear, intimidation, and retaliation," and Hall's efforts to sacrifice "integrity and ethics" for the sake of "test results and public praise" (2017, p .17).

Even today, Atlanta Public Schools has not completely recovered or implemented the type of drastic change required to reform their school system. According to a November 2019 article, 13 Atlanta schools are in the bottom 5% of all schools in Georgia based on student test scores while there are also 13 high schools where fewer than two-thirds of all students graduate (McCray, 2019). Furthermore, while the new superintendent Meria Carstarphen created a program called Target 2021 in order to attempt to identify student victims of the cheating scandal and provide them with additional academic and other forms of support, a Georgia State University study shows scant evidence that the $7.5 million was actually effective (McCray, 2018). These students today are still affected by the deleterious impacts of teachers and administrators prioritizing test scores over learning. One reason for this is that these students may be progressed onto material they

are not yet ready for or have academic supports removed based on their inflated standardized test scores (Sparks, 2016). Again, while these repercussions may be somewhat extreme, they are broadly applicable to other schools and districts. Furthermore, Atlanta represents another case in which the individuals who would actually be rolling out the policies (teachers and principals) were not involved in their creation, creating disconnects and tensions between the district and school levels.

5.3 Conclusion of the Two Case Studies

While it is possible to label both of these case studies as somewhat extreme or anomalous, they are not the only cases where this has occurred and should, at a minimum, give rise to skepticism of a heavily test-dependent system. In both cases, the high-stakes nature of the system and teachers' fear of losing their jobs or suffering severe repercussions led to an unhealthy school and system-wide culture which may have had trickle-down effects for students. As such, even if value-added modeling does yield some tangible benefits, the way in which it is actually implemented may lead to unintended long-term ramifications. Looking back at the framework presented by Hudson, Hunter, and Peckham for successful policy implementation, both of these cases demonstrate profound failures to gather systemwide support/buy-in for either of these accountability systems (2019).

As such, from the case studies, it can be inferred that high-stakes systems based solely on value-added test score modeling as a metric are likely to fail or be unimplementable in practice. In the final section prior to concluding, I look at potential alternatives to value-added modeling that could either replace or supplement it in a more holistic evaluation system. This will ultimately enable me to make some policy recommendations for how to improve the status quo and ensure that teachers are held accountable while not jeopardizing the entire culture of a school or school system, leading to negative student outcomes.

6 Possible Alternatives and Supplements to VAM

In this section, I discuss possible alternative systems to VAM which could either be applied as replacements or supplements to VAM-based systems. These ideas which I briefly discuss stem both from a brief comparative analysis of evaluative frameworks and from concepts generated by education policy experts. While I do not engage in a full-scale comparative analysis of these other systems or generate full policy proposals, I believe that by highlighting these potential options, it can generate possible ideas that can then be tailored to local schools and contexts which might have differing needs for their evaluative frameworks.

In this first portion, I look at a few suggestions by education policy experts before turning to look at alternative teacher evaluation systems in Finland as a drastically different model for teacher evaluation.

6.1 Alternative Systems as Proposed by Education Policy Experts

I first look at proposals for improving and reforming the teacher evaluation system by education policy experts that are not so heavily dependent on value-added modeling or standardized testing.

One such model is Linda Darling- Hammond's report in which she proposes a 5 step model for assessing and supporting teachers (2012). The steps she delineates are as follows:

1. "Start with Standards"—By this Darling-Hammond means a "clear conception of...learning objectives and kinds of instruction...supported by thoughtful curriculum frameworks and materials as well as [meaningful] assessments" (2012, p. 5). This concept enables a common metric for assessing teachers on either neither a national or state level.

2. "Create Performance-Based Assessments"—by which Darling-Hammond means "assessments" that can both "document and help teachers develop greater effectiveness" in their practice

(2012, p .6). Darling-Hammond both cites examples of successful assessments that do this (such as the National Board Certification process) as well as concrete professional dividends that could be tied to each degree of mastery of the practice such as tenure or licensure. While this seems like a reasonable idea in theory, it is imperative that these assessments can actually document outcomes that are beneficial and translatable to students. For instance, there can be a distinction between theoretically knowing how to create a strong lesson plan or different pedagogical techniques and actually being able to implement them in practice.

3. "Build a Standards-Based System of Local Evaluation"—by which Darling-Hammond means tying in the wider system of standards into a localized context in which teachers can both be assessed by "contributions to school-wide goals" and "contributions to growth in student learning" (2012, p .15, 18). In the former category, Darling-Hammond identifies sub-goals involving supporting fellow teachers by sharing curricula and pedagogical practices as well as service to the wider community at large. As to the latter, Darling-Hammond takes a fairly critical view of metrics such as value-added modeling (although concedes that it may have some use if it is clearly connected to the "curriculum" and "appropriateness for the students being taught"). In place of one standardized metric, Darling-Hammond recommends basing evaluations upon a wider array of materials ranging from a "teaching portfolio" (which can show examples of the evolution of student writing or student work) to instructor or school created assessments to student self-evaluations (2012, p .25). Darling-Hammond cites districts such as Rochester, New York and Long Beach, California where some of these materials are taken into account in assessing teachers. One underlying principle in all of these systems is a view of teaching as a collaborative rather than competitive venture where the goal is to ensure success for all teachers and students. As such, there are ways for more experienced teachers to mentor younger ones or for them to take on especially high-need or challenging students.

This conception of education and the potential school or district-wide culture differs starkly from ones that evolve under high-stakes VAM-based systems which can result in a "culture of fear," as highlighted through the case studies. One potential drawback of such a system is how much more challenging it is to evaluate teachers based off of it (rather than a single metric or a few metrics) and to figure out when a teacher has gone very astray. However, it is possible that the potential benefits to students and the school community at large outweigh the difficulties created by those obstacles.

4. "Create Structures to Support High-Quality, Fair, and Effective Evaluation"—through which Darling-Hammond includes structures such as skilled evaluators, supports for teacher assistance, and better principal preparation (2012, p .28). Part of this goal is to enable teachers to have an investment in improving their own practice through collaborative ventures between more and less experienced teachers. Darling-Hammond even recommends use of a "joint committee" with both teachers and administrators to decide on actions such as tenure or improvement plans as well as dismissals (2012, p .33). A program such as this may better meet the goals that Hudson, Hunter, and Peckham highlight through having greater buy-in to the evaluation system rather than an atmosphere of tension or fear between senior leaders and teachers (2019). However, once again, it may be challenging to distill this broad conceptual idea into concrete measures in order to ensure that teachers are given ample and fair warnings and opportunities to improve prior to being dismissed.

5. "Create Aligned Professional Learning Opportunities"—these initiatives should be high quality, sustained (Hammond recommends at least 50 hours over 6 to 12 months), and connected to students and teachers' work through having teachers take on activities such as group planning, reviewing videos of teaching or examples of student work, and peer observations (2012,

p. 33,36). Darling-Hammond also extrapolates this to a broader level through discussing programs that have been implemented in England and Canada through which successful inter-school collaborations have been initiated. Furthermore, Darling-Hammond recommends time is built into the school day for these activities as a core component of teachers' jobs (rather than something that is done as an aside after school or on occasional days). Once again, this is about treating teachers as valued and respected professionals whose input is desired and who have the opportunity to improve their practices to better serve their students.

Articles published by experts at other organizations such as the *ASCD* and *MDRC* agree with and further expand upon Darling-Hammond's conclusions (Toch, 2008; Rosen & Parise, 2017). The ASCD article makes suggestions ranging from a common set of standards to using multiple metrics for evaluation to encouraging teamwork and professional development opportunities, citing The Teacher Advancement Program (TAP) started in 1999 as one such program which implements several of these measures. Potential concerns Toch expresses include monetary costs (which can be around $6250 to $14900 per teacher) as well as some studies which have shown that National Board certified teachers or teachers with high TAP scores don't actually correlate with higher scores on student achievement tests. Toch recommends performance-based pay incentives in order to encourage teachers to engage in this self-improvement. In terms of responding to Toch's concerns, it is firstly important to verify that the program is actually being run successfully by skilled program leaders and with teacher and administrator buy-in and that the components of the program directly align with tangible benefits for students. Secondly, another issue, as discussed earlier, is that there should not just be a unilateral outcome measure by which to assess these programs (as other metrics from emotional intelligence to student behavior also map in meaningful ways onto student life outcomes).

On the other hand, the MDRC article examines professional development in particular (Rosen

& Parise, 2017). In surveys they discovered a gap between principals' views of professional development as being directly informed by teachers and teacher evaluations as compared to teachers who stated that professional development exercises gave them few ways to concretely improve their practice. This issue relates to Darling-Hammond's guidelines in which professional development must be built into and viewed as a core component of teaching rather than a mere supplemental activity, and principals must be given sufficient leadership training in order to be able to implement it successfully.

Ultimately, the advice of several education policy experts is to make teacher evaluation more nuanced with greater attention to students' needs and to create a collaborative, engaged environment in which teachers are treated as professionals with the collective aim of successfully providing high-quality education to students.

6.2 Teacher Evaluations in Other Nations—the Case of Finland

I now turn to a comparative analysis of the US teacher evaluation system to that of Finland. Finland in particular makes for a strong case study comparison location both due to the contrast to the US system and its generally high performance (Walker, 2013). Furthermore, Finland's education system has evolved substantially—moving away from a low achieving, standardized-test based system in the 1970s to one that is more flexible, treats teachers as professionals (and a highly competitive field to pursue), and prioritizes professional development opportunities for teachers (Jehlen, 2010).

Erica Newland as part of a study at Western Michigan University examined the evaluative system in Finland in greater depth (2016). According to Newland, the Finnish evaluation system is a highly localized one through which the National Ministry of Education plays little role. Teachers are highly trained (having to achieve a masters degree at one of the country's highly competitive research university programs) and then are respected as relatively autonomous individuals who

have the ability to develop their own teaching methods and materials. Furthermore, Finland has eliminated standardized testing from their evaluative systems, meaning that evaluations do not rely on standardized tests as a metric.

Instead, teachers are treated as professionals, and evaluation occurs in a "group-based, reflective, and participatory" fashion with the end goal of "creating professional learning communities among teachers and administrators" (Newland, 2016, p .72). As such, Finland eliminated the previous school inspection system and replaced it with a system of trust where teachers are held accountable to their students as well as the wider school community. Furthermore, evaluations aren't seen as punitive but rather as a tool for teachers to further learn and develop their skills. This system has resulted in high levels of success—not only in terms of PISA scores but also in the fact that 90% of Finnish students complete through secondary school with two-thirds of those students then enrolling in universities or polytechnic schools (Newland, 2016, p .70).

The broader point of this comparison is not to say that the US ought to replicate the Finnish system which would be impossible given the different education systems, norms, values, and methods of teacher training. Rather, elements of the Finnish system could be implemented which are in line with Darling-Hammond's (and others') recommendations such as de-emphasizing standardized testing, increasing teacher autonomy, and forming collaborative systems among teachers and administrators. High standards for teachers could remain but be achieved via a broader and more holistic process than standardized test data. Within my conclusion, I examine some of these implications more concretely in terms of possible policy recommendations for reforming teacher evaluation systems in the American context.

7 Conclusions, Policy Implications, and Directions for Future Research

In this final section, I first briefly highlight some of the most salient conclusions from this paper and merge my findings from across the different evaluative lenses—political, quantitative, case studies, and potential alternatives—in considering some of the most essential points when analyzing VAM or other high-stakes teacher accountability systems. Next, while full-scale policy proposals are beyond the scope of this paper and would likely need to be locally tailored, I highlight a few of the possible policy implications of this research in terms of key points that policymakers should take into account when creating teacher evaluation systems. Finally, I recognize that VAM and teacher evaluation are vast and complex issues. While I attempted to be more comprehensive and wide-ranging in discussing VAM than previous work on the subject, this analysis is by no means comprehensive, and there are many other key issues to explore. As such, I highlight some of the limitations of this research and some possible directions for future research on VAM or teacher evaluation more broadly with the US in order to gain further insight into this fundamental topic.

7.1 Conclusion and Policy Recommendations

In summary, there are many reasons to be skeptical of value-added modeling based systems given this holistic analysis. While I don't necessarily endorse eliminating standardized tests or VAM scores entirely, they should be given significantly less weight, and students should be tested less frequently than under the current system. As I discussed throughout the course of this paper, high-stakes VAM has its roots in ideas of students assimilating to certain norms and contributing to American global competitiveness, doesn't fully map onto either tangible or intangible student outcomes, has some unresolved flaws in its quantitative methodology, and frequently fails to generate

significant buy-in from teachers or school administrators. Furthermore, education policy experts have recommended elements of evaluative systems of other nations or certain US districts which have been highly successful through reducing the dependencies on standardized test scores. Other scholars have found that other indices from measures of the extent to which teachers contribute to student behavior to how they affect students' skills like emotional intelligence also map well onto student outcomes.

While it is well beyond the scope of this paper to develop an alternative teacher evaluation system, I can make some more concrete recommendations based on these findings as follows:

- *Reduce the Reliance on VAM and Standardized Testing:* While I do not necessarily recommend eliminating VAM or standardized tests entirely, they should be taken as one minor data point along with others in a more holistic evaluation scheme. Furthermore, policymakers should consider, as Darling-Hammond suggests, how well the standardized tests actually map onto the curriculum that students are learning, the skills that they are acquiring, and the heterogeneity of student learners.

- *Treat Teachers as Collaborators and Autonomous Professionals:* Teachers should be given greater trust and space to collaborate with one another as a core function of their jobs. That collaboration should be oriented towards grappling with the everyday difficulties that they are experiencing and finding mechanisms through which to resolve them. Professional development exercises should be directly related to these goals as well.

- *Develop a Clear Set of Standards for Students and Teachers:* Students should have clear goals which they are meant to achieve differentiated by their current mastery of the material.

- *Use Multiple Methods of Evaluation:* These could range from classroom observations (with experienced other teachers as evaluators) to having teachers create portfolios that illustrate

student work over time to anonymous student evaluations. These measures should be combined holistically in order to assess teachers.

- *Involve Teachers in the Development and Implementation of These Systems:* One way to do this is through the joint committee structure posed by Darling-Hammond where teachers work in conjunction with administrators to develop and oversee these systems. Furthermore, this generates teacher involvement whenever there are concerns about a particular teacher which may result in disciplinary action.

- *Provide Adequate and Ongoing Training for Education Leaders:* Education leaders (such as school principals) should receive ongoing training in how to best develop these systems, conduct classroom observations, and create a positive school environment and culture. Furthermore, school principals should have a collaborative network with other schools in order to address and resolve higher-level problems.

- *Higher Level Education Authorities Should Also Address These Problems in a Broader Light:* Higher education authorities such as local or state Boards of Education or Superintendents should look not just at schools' standardized test scores but rather at a broad range of both quantitative (i.e. graduation rates or postsecondary plans) and qualitative (observations, small group meetings with teachers or education leaders) metrics in order to examine schools' performance. Furthermore, data which has not traditionally been used could also be incorporated ranging from a summary of how students' work has evolved over time to indices of student behavior orm traits such as grit or emotional intelligence. Again, evaluations of when a school or county requires large-scale interventions should be holistic and not tied to any single measure.

While these policies are certainly challenging to implement in terms of the monetary costs and

time required, I believe that they are justified via the substantial benefits to students over the status quo system. It is indubitable that teachers can significantly impact their students' futures both in and beyond the school system and in many ways that go beyond an end of year test score. As such, any metric to assess teachers and any tie-in to either incentives or disciplinary action should be reflective of teachers' overall skills as illustrated by a multitude of different measures and evaluative tools.

7.2 Limitations and Directions for Future Research

While I tried to take a relatively broad approach to this analysis of VAM and VAM-based systems in contrast to much of the current research literature, this review is far from comprehensive, and there are many avenues for the development of future research.

One such research goal would be translating these large ideas into concrete policies on either a school or a district level. Through making these policies concrete and piloting them, issues could be uncovered and addressed in order to see how these evaluative systems actually function in practice. This type of action research study could be useful before creating any kind of broader policy.

Another direction for future research is to further investigate any of the four areas I looked at. For instance, in terms of the quantitative methods component, I examined the debate between Chetty and Rothstein over one issue (random assignment to teachers). However, there are many other quantitative issues at play in VAM which have been analyzed by other scholars. Analogously, there are other case study locations of high-stakes value-added systems and other nations with different models of teacher evaluation. As such, there is room to further expand on any of these ideas or modes of analysis which could help to aid policy development.

Finally, as Darling-Hammond discussed, there are already schools and districts within the US that have begun to implement some of these measures and experienced some preliminary success.

As such, more detailed reviews of those locations could be useful through a variety of methods (quantitative analysis, interview-based work, and ethnography). Once again, the successes as well as obstacles faced within those districts could then be applied in generating concrete and implementable policies which can positively affect student learning and other academic and non-academic outcomes.

References

Aldrich, M. W. (2017, March 20). William Sanders, pioneer of controversial value-added model for judging teachers, dies. *Chalkbeat*. Retrieved from https://www.chalkbeat.org/posts/tn/2017/03/20/william-sanders-pioneer-of-controversial-value-added-model-for-judging-teachers-dies/

Balingit, M. (2018, 24). Racial disparities in school discipline are growing, federal data show. *The Washington Post*. Retrieved from https://www.washingtonpost.com/local/education/racial-disparities-in-school-discipline-are-growing-federal-data-shows/2018/04/24/67b5d2b8-47e4-11e8-827e-190efaf1f1ee_story.html

Boyd, W. L. (1987, March 18). President Reagan's school-reform agenda. *Education Week*. Retrieved from https://www.edweek.org/ew/articles/1987/03/18/2525boyd.h06.html

Breen, A. (2019, December 2). Study: DC Public School's teacher evaluation system continues to improve teacher workforce. *University of Virginia: Curry School of Education and Human Development*. Retrieved from https://curry.virginia.edu/news/study-dc-public-school%E2%80%99s-teacher-evaluation-system-continues-improve-teacher-workforce

Brown, D. (2008, January 14). Senator Kennedy, No Child Left Behind, and the dropout crisis. *The Huffington Post*. Retrieved from https://www.huffpost.com/entry/senator-kennedy-no-child-_b_81354

Chamberlain, G. E. (2013). Predictive effects of teachers and schools on test scores, college attendance, and earnings. *PNAS*, *110*(43), 17176–17182. Retrieved from https://www.pnas.org/content/110/43/17176

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593–2632. Retrieved from https://www.aeaweb.org/articles?id=10.1257/aer.104.9.2593

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*(9), 2633–2679. Retrieved from https://www.aeaweb.org/articles?id=10.1257/aer.104.9.2633

Darling-Hammond, L. (2012). *Creating a comprehensive system for evaluating and supporting effective teaching* (Tech. Rep.).

Dieterle, S. G., Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2013, November 7). *How do principals assign students to teachers? finding evidence in administrative data and the implications for value-added* (Tech. Rep.). Retrieved from https://education.msu.edu/EPC/library/documents/WP30-Dieterle-Guarnino-Reckase-Wooldrige-2013-Student-Teacher-Assignments-and-Value-Added.pdf

Dynarski, M. (2016, December 8). *Teacher observations have been a waste of time and money* (Tech. Rep.). Retrieved from https://www.brookings.edu/research/teacher-observations-have-been-a-waste-of-time-and-money/

Fernández-Berrocal, P., & Ruiz, D. (2008). Emotional intelligence in education. *Electronic Journal of Research in Education Psychology*, *6*(2), 421–436. Retrieved from http://repositorio.ual.es/bitstream/handle/10835/538/Art_15_256_eng.pdf?sequence=1

Flèche, S. (2017). *Teacher quality, test scores and non-cognitive skills: Evidence from primary school teachers in the uk* (Tech. Rep.). Retrieved from http://eprints.lse.ac

.uk/83602/

Fletcher, D. (2009, December 11). Standardized testing. *Time Magazine*. Retrieved from http://content.time.com/time/nation/article/0,8599,1947019,00.html

Hartman, A. (2008). *Education and the Cold War: The battle for the American school*. New York, NY: Palgrave Macmillan.

Hudson, B., Hunter, D., & Peckham, S. (2019, February). Policy failure and the policy-implementation gap: can policy support programs help? *Policy Design and Practice*, *2*(1), 1–14. Retrieved from https://doi.org/10.1080/25741292.2018.1540378

Jackson, C. K. (2018, October). What do test scores miss? the importance of teacher effects on non–test score outcomes. *Journal of Political Economy*, *126*(5), 2072–2107. Retrieved from https://www.journals.uchicago.edu/doi/pdfplus/10.1086/699018

Jackson, K. (2019). The full measure of a teacher. *Education Next*, *19*(1). Retrieved from https://www.educationnext.org/full-measure-of-a-teacher-using-value-added-assess-effects-student-behavior/

Jehlen, A. (2010, October 7). How finland reached the top of the educational rankings. *NEA Today*. Retrieved from http://neatoday.org/2010/10/07/how-finland-reached-the-top-of-the-educational-rankings/

Kennedy, E. M. (2008, January 7). How to fix 'No Child'. *The Washington Post*. Retrieved from https://www.washingtonpost.com/wp-dyn/content/article/2008/01/06/AR2008010601828.html

Klein, A. (2019, January 23). How are states measuring student growth under ESSA? *Education Week*. Retrieved from http://blogs.edweek.org/edweek/campaign-k-12/2019/01/essa-growth-data-state-data-quality-campaign.html

Kurtz, M. D. (2018). Value-added and student growth percentile models: What drives differences in estimated classroom effects? *Statistics and Public Policy*, *5*(1), 1—8. Retrieved from https://doi.org/10.1080/2330443X.2018.1438938

Lawson, J. (2014). Value-added modeling: Challenges for measuring special education teacher quality. *InterActions: UCLA Journal of Education and Information Studies*, *10*(1). Retrieved from https://escholarship.org/uc/item/9r67085n

Mattingly, J. (2018, May 7). 'a Culture of Fear': How the teacher evaluation system Richmond's new superintendent created impacted D.C. schools (and why he won't bring it here). *Richmond Times-Dispatch*. Retrieved from https://www.richmond.com/news/local/a-culture-of-fear-how-the-teacher-evaluation-system-richmond/article_173a63b9-baff-5427-b8e1-b5cc00e9045a.html

McCray, V. (2018, February 22). APS tries to aid cheating victims but impact is small. *AJC*. Retrieved from https://www.ajc.com/news/local-education/aps-tries-aid-cheating-victims-but-impact-small/L3FUfBrQGi7v20JAf1hmCM/

McCray, V. (2019, November 26). 13 atlanta schools among state's bottom 5%. *AJC*. Retrieved from https://www.ajc.com/news/local-education/atlanta-schools-among-state-bottom/OFoehBUux53eIMOtUeRF6M/

Mercury, V., & Hankerson, M. (2019, October 18). Virginia considers deemphasizing test scores in teacher evaluations. *NBC 12*. Retrieved from https://www.nbc12.com/2019/10/18/virginia-considers-deemphasizing-test-scores-teacher-evaluations/

National Research Council and National Academy of Education. (2010). *Getting value out of value-added: Report of a workshop*. Retrieved from https://www.nap.edu/catalog/12820/getting-value-out-of-value-added-report-of-a-workshop

Newland, E. (2016, November). *Case studies of teacher evaluation systems around the world—Chapter 6: Finland* (Tech. Rep.). Retrieved from https://wmich.edu/sites/default/files/attachments/u445/2017/emr_wps2.pdf

Rosen, R., & Parise, L. M. (2017, March). *Using evaluation systems for teacher improvement: Are school districts ready to meet new federal goals?* (Tech. Rep.). Retrieved from https://www.mdrc.org/sites/default/files/iPD_ESSA_Brief_2017.pdf

Rothstein, J. (2016, December 21). *Can value-added models identify teachers' impacts?* (Tech. Rep.). Retrieved from https://irle.berkeley.edu/can-value-added-models-identify-teachers-impacts/

Rudalevige, A. (2003). The politics of no child left behind. *Education Next*, *3*(4). Retrieved from https://www.educationnext.org/the-politics-of-no-child-left-behind/

Shaker, P., & Heilman, E. (2008). *Reclaiming education for democracy: Thinking beyond No Child Left Behind*. New York, NY: Routledge.

Simons, R., & Kindred, N. (2017, 29). *Atlanta schools: Measures to improve performance* (Tech. Rep.).

Sparks, S. D. (2016, April 26). Studies: When educators cheat, students suffer. *Education Week*. Retrieved from https://www.edweek.org/ew/articles/2016/04/27/studies-when-educators-cheat-students-suffer.html

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test peformance of african americans. *Journal of Personality and Social Psychology*, *69*(5), 797–811. Retrieved from http://mrnas.pbworks.com/f/claude%20steele%20stereotype%20threat%201995.pdf

Stein, P. (2019, October 20). Chancellor pledges to review D.C.'s controversial teacher eval-

uation system. *The Washington Post*. Retrieved from https://www.washingtonpost.com/local/education/chancellor-vows-to-review-the-districts-controversial-teacher-evaluation-system/2019/10/20/6c00405c-f0de-11e9-8693-f487e46784aa_story.html

Strauss, V. (2011, May 9). Leading mathematician debunks 'value-added'. *The Washington Post*. Retrieved from https://www.washingtonpost.com/blogs/answer-sheet/post/leading-mathematician-debunks-value-added/2011/05/08/AFb999UG_blog.html

Strauss, V. (2015, December 9). Why it's worth re-reading George W. Bush's 2002 No Child Left Behind speech. *The Washington Post*. Retrieved from https://www.washingtonpost.com/news/answer-sheet/wp/2015/12/09/why-its-worth-re-reading-george-w-bushs-2002-no-child-left-behind-speech/

The National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform* (Tech. Rep.). Retrieved from https://www.edreform.com/wp-content/uploads/2013/02/A_Nation_At_Risk_1983.pdf

The Opportunity Agenda. (2011, October). *Social science literature review: Media representations and impact on the lives of black men and boys* (Tech. Rep.). Retrieved from https://www.racialequitytools.org/resourcefiles/Media-Impact-onLives-of-Black-Men-and-Boys-OppAgenda.pdf

Toch, T. (2008, October). Fixing teacher evaluation. *Educational Leadership*, *66*(2). Retrieved from http://www.ascd.org/publications/educational-leadership/oct08/vol66/num02/Fixing-Teacher-Evaluation.aspx

Walker, T. (2013, March 25). How do high-performing nations evaluate teachers? *NEA Today*. Retrieved from http://neatoday.org/2013/03/25/how-do-high-performing-nations-evaluate-teachers/

Winig, L. (2012, February). *Michelle Rhee's IMPACT on the Washington D.C. Public Schools (Case Number 1958.0)* (Tech. Rep.).