The Value of Value-Added: Science, Technology, and Policy in Educational Evaluation

Daniel Douglas

Graduate Center, City University of New York

Recommended Citation


http://academicworks.cuny.edu/gc_etds/1867
THE VALUE OF VALUE-ADDED: SCIENCE, TECHNOLOGY AND POLICY IN EDUCATIONAL EVALUATION

by

DANIEL DOUGLAS

A dissertation submitted to the Graduate Faculty in Sociology in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2017
The Value of Value-Added: Science, Technology and Policy in Educational Evaluation

by

Daniel Douglas

This manuscript has been read and accepted for the Graduate Faculty in Sociology in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

Date

Paul Attewell
Chair of Examining Committee

Date

Philip Kasinitz
Executive Officer

Supervisory Committee:

Paul Attewell
Stanley Aronowitz
Philip Kasinitz

THE CITY UNIVERSITY OF NEW YORK
ABSTRACT

The Value of Value-Added: Science, Technology and Policy in Educational Evaluation

by

Daniel Douglas

Advisor: Paul Attewell

In the first decade of the 21st century, researchers and policymakers in K-12 education began to focus on evaluating teacher and school performance based on students’ standardized test scores. One evaluative technique, value-added assessment (VAA), has been given particular attention. This research presents a comprehensive study of the theoretical, technical, historical and political dimensions VAA. Theoretically, the assumptions that underlie value-added diverge significantly from the observed operations of the schools and classrooms these models are supposed to evaluate. Technically, even if the theoretical assumptions are accepted, teachers’ actual value-added rankings are shown to be unstable across time periods and classrooms for individual teachers based on publicly-available data from New York City schools. Historical discourse analysis shows how the political and technical evolution of VAA fit a pattern common to prior technical innovations in educational assessment. Finally, making a case study of the Vergara v. California trial, this research demonstrates the political force of VAA data in spite of its known limitations. These findings are considered in the context of sociological theories of science and policy.
ACKNOWLEDGEMENTS

This project challenged and transformed me in ways I could not have imagined at its outset. Overcoming self-doubt, and establishing a sense of place in the academic community, while trying to make an original and interesting contribution to sociology often seemed too complex and overwhelming. My perseverance depended on the help and counsel of many people who will go unnamed here; but it behooves me to mention some of the most important ones. For their support in refining my thinking about education and schooling, I thank my colleagues on the CUNY Academic Momentum Project, particularly Robin Isserles – with whom I shared many conversations on teacher evaluation specifically and K-12 education in general. My broader consideration of science and technology was greatly enhanced by my comrades in a weekly philosophy study group. Specifically, I thank two long term members of the group – Sam Connet and Marzena Marzouk – for their various ideas and suggestions over the years. The members of my dissertation committee, Stanley and Aronowitz and Philip Kasinitz, were infinitely helpful in refining the ideas in this manuscript and in pushing me to find the project’s broader implications. Paul Attewell, my principle advisor, was and continues to be supportive of my scholarship on this topic and others. He is a great colleague, mentor, and friend who has strengthened this work with his substantive and editorial suggestions. My friends and family have provided me with the material and emotional support necessary to make such a long sustained effort. Finally, and far above all others, I thank my wife and best friend Sarah Salman. She has grappled with my ideas as long as I have, always forcing me to think more and more deeply. The life of my ideas in this work depended on her contributions. At the same time, she has been my emotional and spiritual foundation throughout an often turbulent research and writing process. It is impossible to thank her enough for her love and support. Ultimately, even with the many collaborators listed here,
the words are my own; as such, I take sole responsibility for any errors, omissions, or other weaknesses herein.
# Table of Contents

Chapter 1: Introduction................................................................................................................................. 1  
  Accountability, Efficiency, and School Reform .................................................................................................. 6  
  Science, Technology and Policy .......................................................................................................................... 8  

Chapter 2: Value-Added in Theory and Practice................................................................................................. 14  
  The Technology and Politics of Work Measurement.......................................................................................... 15  
  First Difference Models .................................................................................................................................. 22  
  Teacher Quality .................................................................................................................................................. 23  
  A Very Brief Primer on Value Added Assessment............................................................................................... 23  
  Selected Criticisms of Value Added Assessment ................................................................................................. 25  
  Uses, Misuses, and Unintended Consequences of Value Added ......................................................................... 30  
  Sociological trouble with Value-Added: An Authority Paradox ......................................................................... 34  

Chapter 3: Value-Added by the Numbers: A Quantitative Presentation Using NYC Teacher Evaluation Data........................................................................................................ 43  
  Data Source: NYC Value-Added FOIL Data......................................................................................................... 44  
  Findings.................................................................................................................................................................. 47  
    VA “Bouncing:” Temporal Stability of Value-Added Rankings ...................................................................... 47  
    Value-Added “Swaying:” Within-Teacher Variation in VA Rankings............................................................... 54  
    Schools Matter – The Persistence of School-level variables in VA rankings ................................................ 57  

Chapter 4: Value Added in Context: A Technological History of Educational Politics, or a Political History of Educational Technology ........................................................................................................ 63  
  1830-1850: Horace Mann, Samuel Gridley Howe, and the ‘Testing Wars’ of the mid-19th century ................. 64  
  1900-1920: Taylorism, Classroom Management, and the ‘Cult of Efficiency’ ................................................. 70  
  1920-1950: Educational Testing Service, the SAT, and the rise of Mass Educational Testing .......................... 83  

Chapter 5: Value-Added in Private: Developing techniques to measure teacher productivity .................................. 115  
  From Coleman to A Nation At Risk: Economic competitiveness and the narrative of school failure .................. 117  
  Critiques of the Coleman Report and the ‘Education Production Function’ .................................................. 123  
  Early Value-Added Systems: DVAAS and TVAAS ............................................................................................. 125
Clinton, Bush, and NCLB: Education Science and the Focus on Teachers .......................... 131

The Widget Effect, NCTQ, and Political Research on Value-Added ........................................... 138


Background on the Trial .......................................................................................................................... 148

Contextualizing Foundation Influences ................................................................................................. 154

Research Basis of Two Expert Witnesses ............................................................................................. 157

Expert Testimony in Vergara – Raj Chetty .......................................................................................... 172

Expert Witness Testimony in Vergara – Thomas Kane .......................................................................... 183

Expert Testimony – Jesse Rothstein ....................................................................................................... 192

Expert Testimony: Eric Hanushek ........................................................................................................ 197

Outcome and Press Coverage of Vergara .............................................................................................. 202

Chapter 7: Understanding Value-Added: Reason, Rationality and Politics in the Social Sciences ................................................................................................................................. 207

Value-Neutral Science ............................................................................................................................ 208

Science as Value-Laden .......................................................................................................................... 215

On Mathiness and Policy ........................................................................................................................ 228

Bringing it all together – Value-Added and moments of “Methodological Politics” ................. 232

Postscript on Value-Added and the return of politics ........................................................................ 241

Bibliography ............................................................................................................................................. 243
LIST OF FIGURES

Figure 1: Value-Added and the teacher-student-test score chain ........................................... 42
Figure 2: Absolute ranking change between Year 1 and Year 2 ............................................. 51
Figure 3: Absolute ranking change between Year 2 and Year 3 ............................................. 52
Figure 4: Absolute ranking change between Year 1 and Year 3 ............................................. 53
Figure 5: Scatterplot of ELA teachers’ rankings in one year in two separate classrooms ........ 55
Figure 6: Scatterplot of Math teachers’ rankings in one year in two separate classrooms ........ 56

LIST OF TABLES

Table 1: Number of Records for Data Points in the FOIL data file.......................................... 45
Table 2: Merge Results for FOIL and NYC School Progress Report data ................................ 46
Table 3: Ranking movement between year 1 and year 2 – All Math Instructors (N=2,985) ....... 48
Table 4: Ranking Movement between year 2 and year 3 – All Math Instructors (N=2,985) ..... 48
Table 5: Ranking Movement between year 1 and year 3 – All Math Instructors (N=2,985) ..... 49
Table 6: Ranking movement between year 1 and year 2 – All ELA (N=2,795) ....................... 49
Table 7: Ranking Movement between year 2 and year 3 – All ELA (N=2,795) ....................... 50
Table 8: Ranking Movement between year 1 and year 3 – All ELA (N=2,795) ....................... 50
Table 9: ELA teachers’ within-subject, within-year rankings with kappa statistic ................. 57
Table 10: Math teachers’ within-subject, within-year rankings with kappa statistic .............. 57
Table 11: ANOVA tests on change in value added ranking between 2005 and 2006 with school as the grouping variable ................................................................. 58
Table 12: ANOVA tests on change in value added ranking between 2006 and 2007 with school as the grouping variable ................................................................. 58
Table 13: ANOVA tests on absolute change in value added ranking between 2005 and 2007 with school as the grouping variable ................................................................. 59
Table 14: Predicting 2007 teacher VA rank using 2005 teacher VA rank and a school-level variables Sheaf coefficient, OLS Regression ................................................................. 60
Chapter 1: Introduction

As society becomes more dependent on scientific and technical processes to identify and address perceived social problems, it is necessary to better understand how the logic of science and technology interacts with other social logics. The debate over value-added assessment (VAA) of schools and teachers affords an interesting contemporary case precisely because it so prominently features the idea of scientific rigor. In the sense of historian of technology Bertrand Gille (1986), value-added is a technical ‘invention’ that has become an ‘innovation’ – it has proceeded from being a matter of scholarly interest to having consequences for broader social life. Using value-added assessment as a case study, I ask: what is the role of science in the political sphere when scientific method itself is part of the policy?

Teacher assessments based on growth in students’ standardized test scores have become a major point of contention in the long struggle over standards-based public education reform. They were at the center of President Obama’s $4.35B Race to the Top (RttT) program which, beginning in 2009, pushed state governments to compete against one another to change the way schooling is evaluated. The Chicago Teachers Union cited these evaluations as one of the key reasons for a long strike which began and ended in 2012. New York City and Los Angeles both had very tense legal battles over the release of teacher ratings in 2011 and 2010, respectively; the release in Los Angeles reportedly led one teacher to commit suicide. These assessments have been cited as influences in widespread institutional cheating scandals in Washington, DC; El Paso, Texas; and Atlanta, Georgia. More recently, this use of testing has been the driving force behind a successful test boycott by Seattle high school students, a lawsuit against the Florida
department of education by teachers in that state, and a growing opt-out movement staged by parents across the country.

*Value-Added Assessment* (VAA) is the term commonly associated with these measures of teacher quality. Originally imported from agricultural science, value-added assessment was developed as an educational evaluation tool in the early 1990s by William Sanders, then a researcher at the University of Tennessee. Its history can be traced back further to the work of economists, especially Erik Hanushek, who has researched and advocated for outcomes-based assessment of this sort since the beginning of his career in 1968. But it wasn’t until the last decade, with the passage of No Child Left Behind Act and the Education Sciences Reform Act that the value-added system began to be considered as a legitimate tool.

The Tennessee Value Added Assessment System (TVAAS) was the first evaluation program to rely on the education production function concept. TVAAS data, its architects claimed, showed that teacher quality predicts students’ academic growth better than race, socioeconomic status, classroom diversity or class size. It is notable, however, that in the beginning, Sanders was careful to say that value-added data should not be used to make decisions about individual teachers (Hill D., 2000). Other important technical caveats were later made clear by other researchers including statisticians (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004) testing experts (Koretz, 2000) and other economists (Kupermintz, 2003).

---

1 The term Value-Added is used in this proposal as a catch-all to include several different types of assessment which purport to measure teacher ability based on year-to-year growth of student test scores. Other versions, notably Student Growth Percentiles (SGPs) have come into use in some state education systems. I use the term Value-Added to refer to these various methods taken as a whole.

Value-Added assessment did not have a great deal of political currency (or the necessary longitudinal tracking data) in the 1990s and early 2000s – though it was the subject of much academic discussion during those years. Mandatory testing had been implemented for schools receiving Title I funding since the 2001 implementation of No Child Left Behind, and thus provided the necessary test score data. But evaluations still took place only at the school or district level, rather at that of individual teachers, and did not focus on growth measures like VAA. Indeed, before 2009, only four states had any systematic test-based evaluations of individual teachers. But since Race to the Top in 2009, value-added has become a central part of the national school reform debate.

At that time, we also observe a broad shift in the rhetoric around teacher evaluations. A 2009 report by The New Teacher Project, based on surveys of school districts in four states, claimed that the public school system was essentially indifferent to teacher performance except in rare cases when a teacher was marked for removal. The authors called this indifference the “Widget Effect,” referring to the idea that all teachers were seen as equal and interchangeable. Based on

---

3 This shift in rhetoric, which implicitly bemoans the interchangeability of teachers and the similarity of effectiveness ratings, is historically ironic. Curriculum reform movements since the 1960s have attempted to “teacher-proof” classroom experiences – especially those of poor and minority students. The most explicit of these is “direct instruction,” an approach developed and championed by behavioral psychologist Siegfried Engelmann. Direct instruction is conducted by script, affording teachers little if any autonomy in deciding what to teach and when, let alone how. There is a vast literature and a decades-long debate about the efficacy of this method of teaching (National Institute for Direct Instruction, 2011). As with the evaluation reform movements discussed in the work that follows (particularly chapter 3) Direct Instruction was variously resisted in its explicit application, but still has an important impact on American public schooling (Kozol, The Shame of the Nation, 2005). Given this, it is unsurprising that teachers would be found to have similar levels of effectiveness. Value-Added assessment, despite the claim that it leaves teachers free to choose their methods, still results in standardization of instruction by making the test the sole arbiter of quality. Since test questions appear in specific format, and a single approach to each test question is likely to be optimal, instruction in content is reduced to rote delivery of what I would call solution routes. Instead of teaching the concepts
surveys, the study showed that an overwhelming proportion of teachers were given satisfactory ratings, despite teachers’ and administrators’ reported awareness of ‘ineffective’ colleagues. While the report did not specifically call for value-added measures of teacher quality, the differentiations they recommend – distinguishing poor from fair, fair from good, and good from great – are a direct nod to those called for by advocates of test-based assessments (Weisberg, Sexton, Mulhern, & Keeling, 2009, p. 29).

A group of well-known economists (Chetty, Friedman, & Rockoff, The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood, 2011a) went further, making the headline-grabbing claim that differences in elementary school teachers’ effectiveness can have profound effects on students’ adult incomes. Their study, titled “The Long-term impacts of teachers,” claimed that a one standard-deviation increase in teacher value-added during one school year correlated with an increased annual income of $182 per student at age 28 (Chetty, Friedman, & Rockoff, The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood, 2011a, p. 4). They extrapolated this correlation to argue that teacher replacement policies that used value-added assessment could add $1.4 million in lifetime earnings per classroom per grade. Although subsequent research made severe challenges to these claims (Adler, 2013; Rothstein J., 2015), Chetty et al.’s findings were mentioned by President Obama in his 2013 state of the Union Address.

In another politically important moment for VAA, researchers funded by the Bill & Melinda Gates Foundation conducted a massive multi-year study on test-based assessment, called the Measures of Effective Teaching (MET) study. The MET study attempted to randomly assign upon which test questions are based, one teaches the ways to the wanted answer. Teaching to the test is discussed further in chapter one below.
students to teachers – one of the statistical assumptions required for value-added assessment – in order to resolve the question of whether value-added actually measures are invalidated by selection bias. A second component of the study attempted to correlate different types of teacher assessment with value-added measures (Kane, McCaffrey, Miller, & Staiger, 2013). The MET study authors admitted that the experiments were compromised by attrition, non-random sorting of students into classrooms and the voluntary nature of school participation (p.15). A critical review of the MET study further observed that different measures of teacher ability – value-added, peer observation, and student surveys – were only weakly correlated (Rothstein & Mathis, 2013). Both of these findings suggest that the answer to the question asked in the MET report’s title – *Have We Measured Effective Teaching?* – is no.

But from the standpoint of educational policymakers, it seems as if value-added measures are unproblematic. The above reports – but not the corresponding critical responses – gained significant attention. They have been used to argue that 1) current means of evaluating schooling are ineffective, and 2) that teacher quality can be accurately measured by value-added assessment systems. Notable mentions in speeches by former US Secretary of Education Arne Duncan and President Obama, as well as the text of *Race to the Top* itself, have pushed these techniques to the front of the education reform menu.

Calls for test-based accountability also have to be considered against the backdrop of national education policy writ large. Since the Great Recession, lawmakers across the political spectrum have called for austerity measures which have impacted services ranging from social security to NSF funding to mail delivery. In this frame, calls for ‘efficient,’ ‘accountable’ public schooling can be considered part of a strategy for dismantling teachers’ unions in an effort to lower states’ labor costs. While many point out the direct threat that accountability measures pose to teaching
as a profession (Ravitch, 2010) some suggest that these broad-reaching spending cuts constitute a crisis of public values and moral deregulation (Giroux, 2012).

Up until now, Value-Added Assessment has been discussed primarily in terms of its ability (or not) to assess teachers’ and schools’ contribution to student learning. That discussion continues in academic circles, but we nonetheless see a continued roll-out of these policies. Value-added forms of assessment may indeed expand from K-12 to higher education (Chingos, 2013) and to teacher education programs (Guarino, Reckase, & Woolridge, 2012). The task of this project is to understand how and why this happens. Doing so requires treating the phenomenon at the historical and political-economic levels, as well as understanding the social-scientific debate which serves as ammunition in the political battle.

Accountability, Efficiency, and School Reform

The goal of Value-Added Assessment and its predecessors in education is to make schools, teachers, and the whole system of education “more efficient” and “more accountable.” But these words are so often used that they can lose their meanings. Efficient at what? Accountable to whom? Further, the nature of education – or even the narrower concept of ‘teaching and learning’ – is too complex either to augment efficiency or to hold any person or institution accountable by means of quantitative metrics. But it is the evolution of the historical meaning of efficiency in education which motivates this research.

---

4 A few researchers have begun to examine some political aspects of these evaluations; Gabriel and Lester (2013) used discourse analysis of public committee meetings to understand how Value-Added moved from a technology to a practiced policy in Tennessee. Both Baker et. al. (2013) and Pullin (2013) discuss the potential legal challenges that would accompany the use of student test scores in teacher evaluation.
Raymond Callahan’s (1962) *Education and the Cult of Efficiency* looks at the attempt in the early 20th century to bring Taylorism into public schools. Callahan analyzed the rise of ‘scientific management’ in public education, but the policy emphasis was clearly on the latter term in the phrase. Strong management was then seen as the key to school reform; science – in the Taylorist sense – was only intended to inform the changes. For this reason, much of Callahan’s data refers to early classroom management texts by administratively-oriented progressives such as Ellwood Cubberly and John Franklin Bobbitt. Indeed, as historian David Labaree (2010) points out, educational progressives’ key successes in this period were the centralization and professionalization of school administration.

Callahan argues that the vulnerability of school administrators relative to the other political actors – business and political leaders, as well as the media – led to a contradictory attempt to apply the principles of scientific management to schools. What is more sociologically interesting is the book’s assessment of the array of political forces which pushed efficiency and productivity into education, how that agenda was resisted by educators, and the extent to which each side succeeded. At one level, the situation in the early 20th century described by Callahan is much like what we see today; business leaders and elected officials sought to make public education more efficient by asserting control over the way school administration and assessment. Thus, one aspect of this project is to update Callahan’s story in terms of existing political forces.

Diane Ravitch’s (2010) broad treatment of current school reform efforts, including test-based assessment, identifies at least some of the principle political actors in education reform and some of the ideological assumptions that motivate them. She argues that policy is being driven by a coalition of lawmakers concerned with ‘educational standards’ and private foundations interested in data-driven decision-making and the privatization of public schooling. In her work on the
history of No Child Left Behind, DeBray (2006) suggests that education policy has been a particular focus of the growing efforts of conservative think-tanks since the mid-1990s. 5 But these accounts leave unexamined the roles of politicians and social scientists (of all political leanings) in the process of creating and enacting this particular version of assessment.6

Science, Technology and Policy

But the technologies of assessment have also clearly evolved since the period described in The Cult of Efficiency, as has the perception of technology’s role in education. The capacity to collect and analyze numeric data at unprecedented levels of complexity places education research and statistics in a much more central place than was the case at the turn of the 20th century.

Accountability and efficiency have been “buzzwords” in public school reform for as long as there have been public schools to reform. But the growing capacity to quantify the process of teaching and learning – or the belief that such a capacity exists – changes the meaning of these terms. Political commentator Evgeny Morozov (2013) makes the broad point that as technology advances, policymakers increasingly rely on ‘quick fixes’ afforded by computers, risking what

5 Elizabeth DeBray’s Politics, Ideology and Education (2006) notes that the accountability architecture for NCLB comes from a cadre of “New Democrats” and that the Bush Administration passed this legislation against the opposition of Congressional Republicans.

6 The role of the teachers’ unions in shaping evaluation policy is certainly as important, if not more so, than that of social scientists. Given its implications for school labor practices, whatever its impacts on student learning, the Obama administration’s ‘Race to the Top’ is for teachers a ‘race to the bottom.’ Brill (2010) has portrayed these evaluation policies as ‘The teachers union’s last stand.’ Thus, it is critical to analyze how unions and their leaders have made sense of and either struggled against or embraced these policies. Stanley Aronowitz (2014) examined the different variations of organized labor’s response to technical innovations as ‘adaptive’ or ‘interventionist’ strategies. The former approach accepts technological change as inevitable and seeks to bargain for the existing membership, come what may in the future. The latter seeks an active role in determining whether and how technology will be implemented in the workplace. The role of teachers’ unions is considered at various points in the research below, but a thorough examination of this aspect of education policy creation is beyond the scope of this study.
he calls “the folly of technological solutionism.” The political dimension of education reform must be understood in light of these technical changes.

Thus, while Callahan’s research questions how ‘scientific’ principles inform the management of schooling in the early 20th century, in the contemporary situation, science seems at least as important to understand as management (if not more so). The relationship between science and policy has been theorized following two broad themes.

One broad position regards science as a value-free enterprise. This perspective on science can be traced back as far as Plato, but is first brought to the social sciences in the work of August Comte (1853/1998) and Emile Durkheim (1895/1982) under the heading of positivism – the position that scientific method can provide us with unmediated access to truth. While positivism has been widely criticized and updated in the field of philosophy (Phillips & Burbules, 2000), it continues to constitute the implicit assumption of much social scientific work. In this frame, scientists stand apart from political interest provided they closely adhere to methodological principles, and science itself is only politicized when the knowledge it generates is coopted by interested parties (for example see Henig, 2008).

Theorists in more critical traditions argue on the other hand that science contains its own implicit values. The Frankfurt school views scientific reason and the technology it produces as reinforcing a logic domination which is connected to the commodity form posited by Marx (Horkheimer & Adorno, Dialectic of Enlightenment, 1944/2002; Marcuse, One-Dimensional

---

7 Reese (2013) also examines the question of testing and its impact on educational reform, but begins his investigation in the mid-19th century and focuses on Massachusetts. He too suggests that the primary motivation of progressive reformers was to change the structure of schooling, but notes that the emphasis on written tests was animated by a fascination with the nascent science of statistics.
Man, 1964/1991). French social theorists, under the broad heading of post-structuralism, emphasize the socially constructed nature of scientific knowledge and view science as a dominant regime of truth which consequently shapes the consciousness of both scientists and non-scientists (Foucault, The Order of Things, 1970/1994; Althusser, 1990/2012). Still another strand of critical theory drawing from diverse perspectives focuses on scientific method as the mechanism that allows science to become dominant over other forms of discourse (Feyerabend, Against Method, 1993). Theorists in this tradition propose that challenges to this dominance must disrupt the pervasive belief in value-free, objective science and return science to the field of discourse (Aronowitz, 1988).

Proceeding from these theoretical perspectives, and taking value-added assessment as a contemporary case study, this research asks two basic questions: How does a technique like value-added, initially conceived as a tool of academic inquiry, become a policy tool? Further, once this transformation takes place, what is the role of science and scientists in the public sphere when their techniques become policy tools with social consequences? I address these questions in the following chapters.

Chapter one looks at value-added assessment in theory and practice and thereby summarizes the technique, its basic justification, and some of the main threads of technical criticism. The questions addressed here are: how do the theoretical assumptions of value added modeling in its various incarnations align (or not) with the ways in which schooling actually happens? Further, what might a schooling system that aligns with such a technique look like? I argue that key VAA assumptions about classrooms and schools – random assignment of students to classrooms, stable student retention, annual pre- and post-testing, and most importantly attribution of students score gains solely to teacher quality – vastly misrepresent what actually happens in schools.
Further, relying on Willard Waller’s (1932/2014) *Sociology of Teaching*, I contend that the emphasis on test-taking required by VAA presents serious challenges for both classroom management and for the larger moral project of schooling.

Chapter two examines New York City VAA data and demonstrates some major problems with the technique. The data for New York City were made publicly available by a 2012 freedom of information act (FOIA) request by some of the city’s major news outlets. They have since been the subject of a few journalistic treatments, but none have merged the data with school-level characteristics, which are also publicly available and are included in my analysis. The question answered here is: how do value-added assessments actually behave in practice? Analysis of these data first reveals that VA measures fluctuate from one year to the next. Next, relying on a subset of teachers who teach the same subject to more than one group of students in a single year, I find that even within the same year and subject a single teacher’s rating can be vastly different between classrooms – suggesting that VA is not a stable trait within individual teachers. Finally, I use regression analysis to show that teachers’ VA scores are significantly impacted by certain school characteristics although VA models claim to have adjusted for school-level influences.

Chapter three investigates how the history of evaluations research and education policy led to the practice of evaluating teachers on the basis of students’ test scores. This chapter is a synthetic treatment of existing work on both social science methodology and on the history of ‘scientific’ education policy. I analyze the early genesis of paper-based testing in Massachusetts in the mid-19th century, the push for scientific management of education documented by Callahan (1962), the rise of standardized achievement testing between World War I and the 1950s, and the expansion of systems analysis techniques into social and education policy in the 1960s and 70s. This historical treatment of educational assessment innovation shows that repeatedly when new
evaluation technologies are introduced into the field of education, they appear as both ‘panic’ and ‘panacea.’ They are first used to diagnose familiar problems of schooling, albeit at new levels of detail, and with increased vigor. They are also accompanied by unreasonable expectations for solving supposed education problems. Ultimately, however, the enthusiasm for these technological fixes wanes either due to political resistance by educators, or when new techniques supersede them.

Chapter four takes the social science literature on value-added itself as an object of a discourse analysis. The science behind value-added has been mustered as evidence by critics and proponents alike. This chapter answers the following question: how does the discourse around a technique like value added evolve as it moves from a matter of academic interest in the invention stage to one of political relevance in its innovation stage? I treat the scholarly work on VAA in two phases: first from 1994 until roughly 2008, when discussion of using student test score gains to evaluate teachers was primarily occurring within academic circles; then, from 2009 to the present, when the federal Race to the Top Program thrust VAA into the forefront of policy development. This analysis shows that as VAA transformed from a research method to a policy tool, critical discourse about its fitness for the task of teacher evaluation was drown out by more zealous advocacy of VA as a policy tool.

Chapter five seeks to clarify the use of scientific knowledge of Value Added Assessment in recent struggles over its implementation as policy. I do this by analyzing the 2014 Vergara v. California case, which made value-added assessment the centerpiece of expert testimony in a lawsuit which sought to overturn teacher tenure protections in that state. This chapter explores how value-added research was understood in a courtroom setting and, further, how is the ideal of science used in pursuit of political objectives. I find that expert testimony by proponents of VAA
understated the dubious assumptions required for modeling teacher effectiveness in this way. Further, when opposing experts sought to explain and critique these assumptions, they were regarded as less scientifically rigorous than those who presented ‘objective data’ from VAA research.

The sixth and final chapter brings these findings together and attempts square them with broader perspectives on science and technology articulated in social theory. Specifically, this chapter traces the literature on positivism in the social sciences and critical theories which argue that science is constituent of power in modern society. I conclude by arguing that the dominant belief in science as a value-free enterprise has peculiar consequences when the tools of research become tools of policy.
Chapter 2: Value-Added in Theory and Practice

On February 24th 2012, the New York City Department of Education released value-added performance rankings of 18,000 teachers in schools across the city’s five boroughs, which were subsequently published (in some form) in all the city’s major newspapers. This came after a lawsuit by the United Federation of Teachers failed to prevent the release of the data on grounds of privacy. In the days and weeks following, commentary came flowing in from corporate philanthropists (Gates, 2012), city officials (Walcott, 2012), and well-known education researchers (Hanushek E., 2012; Darling-Hammond, 2012). For most of these commentators, what was at issue was not whether or not there should be a system in place to evaluate teachers, but the public nature of the data release and the uses to which the data will be put.

New York was not the first school system to make its teacher ratings public. In April of 2011, the Los Angeles Unified School District released value-added rankings of schools, and followed that with the release of teacher value-added scores in May. This release was prompted by the Los Angeles Times conducting their own value-added assessment in August of 2010 and publishing teachers’ names and rankings, an event that reportedly led one teacher to commit suicide after receiving a low ranking (Zavis & Barboza, 2010).

While the public release of rankings was notable in these two major cities, the use of Value-Added Assessment (VA) has spread to school systems across the country. Indeed, it has become the preferred method of teacher evaluation for states interested in the federal “race to the top” program which mandates that districts competing to receive funding tie teacher evaluation to
student performance. Further, many states and school systems have begun using these teacher assessments to make decisions on promotion, tenure, compensation and dismissal of teachers.\textsuperscript{8}

Since the development of the first functional value-added assessment system in Tennessee in the early 1990s, economists, sociologists, statisticians and policy analysts of all stripes have debated the validity and proper use of this approach to assessing school and teacher performance. While this chapter intends to make some contribution to those debates, the larger purpose is to put value added assessments in an historical context of education policy and assessment technology. We begin by looking at some theories of work measurement; then we will consider the technique and political roots of value added assessments in this context. Next, we will compare the theoretical assumptions of VA to the reality of its application in schools, thereby highlighting some of the methodological criticisms. Finally, we look at the use, misuse and potentially hazardous effects of these assessments to the extent that they resemble and differ from those of other forms of workplace accountability.

The Technology and Politics of Work Measurement

While Value-Added models are a relatively new development in the history of education, the basic principle of measuring work by outcomes is really quite old, perhaps as old as economics itself. Sociologists often consider work measurement through the lenses offered Max Weber and Karl Marx.

Max Weber’s concern with rationalization was a macro-historical argument about a shift in worldview and belief (one that replaced non-rational explanations of natural and social

\textsuperscript{8} By 2014, 45 of the 50 states were on track to have the necessary data systems for VA assessment (National Council on Teacher Quality, 2014).
phenomena with a rational or scientific attitude towards cause and effect, means and ends) and simultaneously an argument about the emergence of new institutional forms. Bureaucracy enhanced top-down authority, improved information transparency and ‘accountability,’ and facilitated rational means/ends decision-making. According to Weber, written records, rules and standardized operating procedures, defined responsibilities and duties within a hierarchy, and employment of credentialed experts, were all mechanisms that made administration more predictable, accountable, and enduring, and therefore more effective than earlier forms of social organization. Simultaneously, Weber highlighted the development of tools for rational organizational decision-making, in particular emphasizing capital accounting and double-entry bookkeeping (Carruthers & Espeland, 1991).

Scholars have followed in Weber’s footsteps by documenting the evolution of managerial and organizational technologies that enhance information flow and decision-making. Chandler (1977) and Yates (1993) examined the rise of professional management in the US, identified shifts in organizational forms and – in Yates’ case – new tools for organizational communication. Drucker (1954) pioneered ‘management by objectives’- a technique for aligning organizational activities with desired goals. World War Two saw the development of statistical quality-control techniques that also increased predictability and uniformity in production (Deming W. E., 1982). Much of modern management theory can also be viewed as efforts to improve the rationality of planning, decision-making and control within bureaucratic organizations, through a variety of structural and procedural innovations (Micklethwait & Wooldridge, 1996).
From a Weberian perspective, the movement towards value-added assessment of teachers and schools can be read as a recent example of a centuries-old effort to extend bureaucratic methods of accountability into previously exempt settings, a small step forward in the rationalization of education. Value-added assessment might prove as important a new tool in education as double-entry bookkeeping once was for industry. What the Weberian perspective does not consider, however, is that efforts at increasing managerial control may go well beyond the rational use of information for decision-making.

Another tradition in industrial sociology draws its inspiration from Marx, and emphasizes managerial control over employees rather than rationalization. Marx noted that employers purchase the labor power of employees, their potential for work. It therefore falls upon management to ensure that this potential is actually transformed into work done, into marketable products and services – for capitalism to survive, exploitation must proceed. Since this perspective views workers’ employment in capitalist firms as necessarily alienated, a struggle ensues between management and labor over the degree of effort expended, the pace and intensity of work, and the degree of autonomy and self-direction of workers (Bendix, 1954). Scholars in this tradition have documented different types of control over employees, and different forms of resistance and/or acquiescence that vary either historically or across different occupational niches in the economy (Burawoy, 1979; Edwards, 1979).

Measuring the work done by individual employees and paying them according to measured output is one of several managerial control strategies that have waxed and waned over time. In the early 20th century, Frederick Taylor (1911) championed a strategy called “scientific management” that involved studying manufacturing jobs in order to discover the most efficient
methods to (say) machine a particular metal component. Workers were then shown the “one best way” of accomplishing their task; their output was measured, and they were incentivized to follow that behavioral recipe through piece-rate and related payment-by-results schemes. The Gilbreths applied similar techniques to routinized office work.

Taylorism had unanticipated consequences that blunted its spread. It intensified industrial conflict by making the struggle between labor and management very visible: setting the piece rate affected workers’ ability to earn. Management frequently cut rates when they discovered employees earning too much, a response that Taylor himself viewed as irrational because it focused on managerial prerogative and status distinctions rather than the maximization of output. Workers then developed various strategies for subverting work measurement, including output restriction. Worker turnover, unionization, and strikes increased (Aitken, 1960). In practice, Taylorism intensified struggles over control at the workplace, undercutting tacit effort bargains that tended to emerge between employees and management.

Computerization of clerical work in the 1960s and 1970s led to predictions that computer surveillance would lead to a revival of work-measurement, piece-rates and payment-by-results. Attewell (1987) reviewed the evidence and concluded that this development was unlikely except in jobs that were highly routine and required little employee discretion, and where employees were in such abundant supply that high turnover rates (and possibly unionization) were acceptable to managers. Today however, to the degree that so-called “Big Data” can be used to document professional work like that of doctors and teachers, it seems possible that this could change.

---

9 Scientific management is considered in great detail in Chapter 3.
In the past, payment-by-results has not been common in non-routinized and professional jobs, except situations where there is a clear numerical measure of individual output. Some salespeople are paid commissions or bonuses based on their sales volume, and some lawyers receive bonuses based on billable hours, but these are exceptions to the more common situation, where professionals are paid a salary, and where individual work measurement is not prominent.\footnote{Attempts to reward teachers and other professionals are also considered in Chapter 3.}

One personnel practice that emerged and spread since the 1960s is the practice of having an annual review of a professional or white-collar employee’s performance, usually undertaken during a meeting between the employer and a supervisor. One variant, associated with large firms like IBM and General Electric involved assigning a letter grade to the performance of white collar employees, based on the qualitative judgment of the supervisor, rather than using objective work measurement. In its most punitive versions, writers suggested that these grades would be curved, so that (for example) the bottom quarter of employees who received a low grade in the annual assessment would be fired (Deming W. E., 1982, pp. 101-106). There is no evidence that this punitive version of performance evaluation was widely adopted; it seems unlikely that large proportions of employees would have been laid off annually based on such reviews without this being noted by the business media. Moreover, the expense of recruiting and retraining one quarter of employees, not to mention the intensification of conflict in the workplace, would be considerable.

A related development, facilitated by computerization of transactions and records was a proliferation of sales and production “targets,” creating multiple performance measures.
Employees could then be held responsible for whether they were reaching their targets on several of these measures. W. Edwards Deming, the father of quality control, was highly critical of this kind of assessment because it pressurized employees to improve performance without giving them any understanding of how to better meet those targets. It held individuals responsible for outcomes over which they may have had limited control. As he put it: “…management by numerical goal is an attempt to manage without knowledge of what to do, and in fact is usually management by fear” (Deming W. E., 1982, p. 76).

Linking assessment and/or pay of teachers to value-added measures of student performance is in one sense an extension of this type of control via targets. But it is novel insofar as it seeks to measure and reward the quality of a professional’s work performance, rather than measuring effort or output indicators like numbers of children taught. And it aims to do so through “objective” quantifiable measurement, rather than using a supervisor’s judgments. Since this application of value-added measures is in its infancy, and current systems are linked to modest bonuses rather than large scale layoffs, it is hard to judge where this new form of workplace control might go. However, if we draw an analogy to the findings of industrial sociology of the past, we would expect that using value-added measures as a form of “high stakes” assessment of teachers with substantial penalties and rewards would lead to gaming and cheating on assessment measures, cause increased tension and labor conflict within schools, and could thrive only in contexts with a plentiful supply of new teachers to replace those leaving.

Evidence has already emerged of widespread cheating on high stakes tests in schools, carried out by teachers and principals, even before a value-added assessment has been deployed for assessing individual teachers (Vogell, Perry, Judd, & Pell, 2012; Boyette, 2011). In the current economic climate, there may be a steady supply of new teachers seeking jobs. But there is
already a high level of voluntary teacher turnover, especially in schools that serve low-income students, where recruitment is sufficiently problematic that many staff members teach outside their areas of specialization (Ingersoll, 2003). Moreover, projections of teacher retirements have led commentators to predict looming teacher shortages.\footnote{Jesse Rothstein (2012) discusses the implications of value-added measures of teacher effectiveness in the context of teacher labor force shortages. That paper and other statements related to teacher labor force issues are discussed in chapter five below.} Based on past research on workplace control, these features would limit a control regime in which teachers were strongly rewarded or sanctioned according to value-added assessments of individual teachers.\footnote{There are also diverging views on whether turnover is a problem. Teaching is a “back-loaded” profession, in which many of the most enticing benefits come at the end of one’s career – salary increases and often generous pensions. Professional autonomy also increases for veteran teachers who are afforded tenure protections (also see chapter 5 below for more on this issue). Given these structural factors, many education reformers see high teacher turnover as a feature, rather than an unintended consequence, of heavy-handed evaluation policies – as they removes teachers before they become expensive and relatively hard to control. Gordon Kane, and Staiger (2006) argue that ‘reducing barriers to entering teaching’ should accompany the introduction of test-based evaluation systems. A similar impulse animates programs like Teach For America, which has eager recent college graduates enter the classroom for a short time (usually two to three years) with the expectation that they will move on rather than become career educators.} Based on past research on workplace control, these features would limit a control regime in which teachers were strongly rewarded or sanctioned according to value-added assessments of individual teachers.

The immediate problem for doing work measurement of teachers is: how does one measure the work? Unlike industrial labor and some clerical tasks, teaching can’t be rated simply according to speed or accuracy of execution at the level of the process. Moreover, the hurdle is raised even further if one wants to be ‘objective’ and measure that work based on outcomes rather than on teacher performance. The outcome of teaching – educated students – is not at first glance something inherently measurable. Thus, to argue that teachers’ work can be measured in the way of other work processes, two major uncertainties must either be overcome or avoided: first, that some measurable student outcome faithfully represents her educational achievement; second,
that some part of this achievement can be precisely and correctly attributed to the work done by an individual teacher.

First Difference Models

At the most basic level, value-added assessment relies on a technique called “first difference models. 13” For educational assessment, the ‘difference’ in question is between a pretest score and a posttest score on some measure of student achievement. Between the two tests is the intervention – teaching – and the difference between the two scores is said to be the effect of that intervention. James Coleman and his colleagues (1966) utilized first difference models to examine the impact of school segregation on the academic outcomes of minority students and later on the differences between public and private schools (Coleman et al 1982). This work stands at the beginning of a long line of research that has used some form of value-added assessment, but at the level of the school rather than of the individual teacher. The difference between evaluating an individual and an aggregate (like a school) is of both practical and theoretical consequence, which will be addressed in the following sections. Nonetheless, in theory, by using a first difference model, Value Added Assessment is able to ‘objectively’ – that is based on outcomes rather than on ‘subjective’ observation of teaching – assess school or teacher effects on student learning growth.

13 Methodologically, Allison (1990) looks specifically at the use of first differences or “change-scores” as dependent variables in regression models, and outlines the use of first differences in ‘fixed effect’ models with longitudinal data (Allison 2009). Liker and his coauthors (1985) offer a good comparison between two wave panel data – which looks at different cohorts over time to measure trends – and first differences, which look at the same cohort over time.
Teacher Quality

Given that a tool exists to measure – however accurately – the ability of teachers to affect students’ achievement, the data collection needed to make such evaluation possible is both costly and complicated. Furthermore, attempting to impose such an intrusive measurement system is likely to generate resistance, as it has in other workplaces. Thus, some amount of political will and intellectual justification must be present to initiate the process. Economist Eric Hanushek has advocated for identifying effective teachers based on student achievement since the 1970s. Other economists have recently made this case drawing on larger data sources and newer more sophisticated methods (Gordon, Kane, & Staiger, Identifying Effective Teachers Using Performance on the Job, 2006; Rockoff, 2004). The common thread among these studies is that they seek to place an effective teacher – measured by her ability to produce high score gains on standardized tests – as the centerpiece of an economically efficient classroom while simultaneously challenging assertions about the positive effects of decreasing class size (Milesi & Gamoran, 2006), better teacher training and credentialing (Ballou & Podgursky, Reforming Teacher Preparation and Licensing: What is the Evidence?, 1997; Aaronson, Barrow, & Sandler, 2003), and more generally increasing school funding (Hanushek E., 1997).

A Very Brief Primer on Value Added Assessment

Value added assessments are said to measure teacher effectiveness based on student achievement growth. In the past, school performance has been evaluated by changes in grade-level test scores from one cohort of students to the next – comparing this year’s third graders to last year’s. Thoughtful supporters of value-added systems criticize this older system as unfair (Harris D. N.,

---

14 Hanushek’s contributions to Value-Added will be considered in greater detail in Chapter 4.
2011), as it penalizes schools – and teachers – for having students who start out with lower test scores and thus will always be seen as lagging behind despite efforts made to improve. As a growth model, value added attempts to account for where students start, measures changes in those students, and judges progress and effectiveness of instruction on that basis.

How value-added assessment works in theory is quite simple. Here, we imagine a single teacher, in a single-grade classroom, teaching only one subject to a group of students. Even this last point can muddy the terrain; we make a serious leap by saying that any subject can be taught without reliance on the others. But, that is one of many discussions for later; for now, we are speaking about – and perhaps clarifying – the ideal conditions under which this form of evaluation would be conducted.

To measure value-added for this teacher, one needs two tests: a pretest which is administered before the teacher works with the class, and a posttest which is administered after that teacher has finished working with the class. Each student has a score on both tests, and the difference between each student’s two scores – posttest minus pretest – is that student’s “gainscore.” Then, we average the gainscores of all the students in the teacher’s class, and we have an average gainscore. This is the measure of what the teacher contributed, by his instruction, to the class’ learning in that subject in that timeframe. But the average value-added, on its own, does not tell one very much about the quality of that teacher, as it’s just a single point.

Once this average is taken, it is compared to other teachers who teach the same grade and subject to their classes. Each of them also has an average gainscore for the year. With enough teachers and classes, the data points begin to form a distribution of average gainscores. The distribution of average gainscores has a mean, which is said to be the effect of the “average teacher” on student
achievement growth in that grade and subject. Our original teacher would then be compared to the average teacher. Based on this, he could fall into three categories: below average, average, and above average; using the same system, our teacher could also be assigned a percentile ranking. In this way, value-added becomes a relative measure. This is important because in such a relative system, there will always be some group of people performing below the average level of effectiveness.

To summarize: we now have, based on pretest and posttest measures for an entire class (or classes), an estimate of an effect for our teacher; we have used it to compare our teacher to others teaching the same subject at the same grade level.15

Selected Criticisms of Value Added Assessment

Having established the basic principles of how value-added assessment utilizes a first difference model to attempt an ‘objective’ measurement of teacher effectiveness based on student outcomes, we now move to consider some of the salient criticisms of value-added assessment as a technical process. What all the criticisms in this section share is that they address the mismatch between the theoretical principles of value-added and the realities of the work to which they are applied. For each criticism mentioned, we will address any modifications that have been made in practice or have been suggested by the research. For some of the criticisms, we also share our own correctives which we have not seen addressed in the literature.

Value Added Models are ideally based on random assignment of students into classes, but students are almost never randomly placed in either classes or schools (Kupermintz, 2003; 15)

15 In the Appendix to this chapter, I provide a diagram which shows the many points of interaction which Value-Added leaps over in inferring teacher quality from student test scores.
Rothstein J., Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement, 2010)—and in most school systems in the country, precisely the opposite is true. In large school systems, students are placed in schools based on both residence and preference. Within schools, there is usually some degree of tracking into general education, gifted programs, and various forms of developmental education, which is to say nothing of the various programs for which students are pulled out of their ‘regular’ classes during the day. Now, let’s imagine a gifted classroom to see a different angle. The students in this class have likely scored high on the pretest and will likely do well on the posttest, demonstrating achievement growth. But how much of that gain is attributable to the teacher and how much to the relatively more diligent students he teaches and to their families? Further, what ought to be said about peer group effects? Here, the issue of separating measured teacher effects from actual teacher effectiveness may end up giving the teacher too much credit. It is thus little surprise that in New York City after the release of teacher VA ratings, the Daily News was able to print a subhead which read “Surprise: the best teachers are at the best schools” (Sandoval, Lesser, & Chapman, 2012). An analogy to factory work would be to compare the efficiency of two groups of sheet metal workers, giving one group pre-measured sheets and the other not, and evaluating the teams as if they were equal on all fronts. This is not to say that Value Added treats everyone equal and makes no attempt to adjust statistically for the non-random assignment of students into classes; it is simply that given the degree of non-randomness that exists in most school settings, statistical adjustments may not be enough.

One variant of non-random distribution of students is that certain variables—for example race and socioeconomic status—have long been linked in sociological and educational research to lower performance on tests. If teachers are to be held to account for achievement growth, it
seems important that these characteristics should be included in a quantitative assessment of a teacher’s ability based on test scores. While the value-added system developed by Sanders uses a pure first-difference model which assumes that controlling for a students’ ability in the pretest is enough to obviate the inclusion of student characteristics, another system developed for the Dallas schools primarily by William Webster and Robert Mendro (1995; 1997) utilizes Hierarchical Linear Modeling (HLM) and multiple statistical controls for student characteristics in addition to controlling for prior achievement. The fairness variable model originally used by the Dallas schools has gained favor in some municipalities such as New York City, while the original Tennessee system has been implemented in others including North Carolina.

The inclusion of ‘fairness variables’ as statistical controls in the modeling of teacher value-added is a clear response to the problem of non-random distribution. But the distribution of student problems is not fully accounted for by this response if we imagine student learning following a model other than a simple linear process. Specifically, value-added models assume that student achievement – as reflected by test scores – grows along a relatively straight line and at a consistent rate from one year to the next. The changes in time (x) are independent of the changes in growth (y). But it is just as possible that learning happens in such a way that over time the rate of learning growth changes. Our hypothetical model follows the notion compound interest – in which you collect interest that becomes part of the principle and subsequently accrues interest of its own. Simply, as students learn more, the “more” that they have learned further enables them to learn more in the next time interval. This may pose no problem if these students are randomly distributed in classes, but when they are not as is typically the case, we could think of a class of higher ability students as initially able to accrue more knowledge and thus able to build on this in subsequent years. If student learning were to follow this model, measures of teacher
effectiveness based on student learning gains would have to account not only for the way certain variables affect the rate of growth, but also the way they affect the growth of the rate.

Having outlined some concerns related to students, we now move to the source of the measurements – the exams taken by students. We hold aside the debate about whether testing in general is an accurate measure of what students know or learn and consider whether or not the tests being used are designed the task of evaluating the effectiveness of teachers.

As noted above, measuring value-added depends on successive testing of students both before and after the intervention of a teacher to measure the teacher’s effect and relative effectiveness. As such, these tests should – following the norms of sound scientific practice – occur immediately before and immediately after the intervention takes place. But, as with the distribution of students into classrooms, actual school practice varies significantly from this evaluative ideal. Students do not typically take a test on or before the first day of school to determine what they know at “t=0.” Further, they do not typically take the tests being used in value added assessments at the very end of the year. Since these tests are often used to evaluate student progress and are used in student promotion decisions, they need to be graded before the end of the year. Moreover, if there is only one test per year, then the prior year’s post-test becomes the pre-test for the following year’s teacher. Two obvious problems emerge. First, two teachers’ work is being measured with each test – the end of “last year” and the greater part of “this year” in question – but only one “this year’s” teacher being evaluated by it. Secondly, the test does not account for the intervening summer. So, our fourth grade teacher will be evaluated partly on the performance of her students’ third grade teacher(s) and will be rewarded or penalized for whatever her students did or did not retain from the previous year after the summer. While examining the effect of different test instruments, Papay (2011) finds that test
timing does have an effect on teacher value added estimates, more so than the difference he finds from the use of different standardized test formats. Summer learning loss has been studied and found to be unequal across different groups of students and in the expected directions (Downey, von Hippel, & Broh, 2004; Alexander, Entwisle, & Olson, 2001). McCaffrey and his colleagues’ (2004) very carefully argued work suggests that the assumption of VAA that teacher effects persist undiminished is a significant flaw in the model specification.16 Finally, as we shall see in the following section, the work of the prior year’s teacher can have unexpected consequences.

Many authors have argued that, though there is enough data within an entire school to generate reliable estimates of the school’s value added, using the same information for individual teachers presents an issue of statistical power. Most teachers don’t teach enough students, and therefore don’t have enough students’ test scores, to generate reliable estimates of their value-added. This is especially true of teachers in their first few years and for teachers who teach smaller classes (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Braun, 2005). Acknowledging this problem, the Tennessee system does not evaluate special education classrooms, where the classes are too small to produce reliable estimates (Topping & Sanders, 2000). Since these measures are intended in many systems to determine retention and promotion of new teachers, the problem of statistical power cannot be put aside, as it is precisely the individuals who are most affected whom these systems do the poorest job assessing. If a teacher is considered for tenure after two years of teaching, and promotion decisions must include this type of assessment, annual test score changes of a few dozen students will partially determine that teacher’s career for years to come.

16 See also Briggs and Weeks (2011).
The problem of missing data is particularly acute for teachers in urban schools, and especially among low-SES students, where moving from one school to another is a regular occurrence. If a student moves between schools, then that student’s change in test scores – good or bad – cannot be meaningfully attributed to any one of the teachers she has interacted with. Further, at the level of the school, fluctuating populations could make a school’s measured effectiveness somewhat meaningless, as they don’t serve a stable population (Baker, et al., 2010; Corcoran, 2010).

Uses, Misuses, and Unintended Consequences of Value Added

So far, we have discussed both what value-added intends to do in theory, and how that ideal squares with the realities in the schools where such assessments are used. The general point is that while value added purports to measure what students learn in school, it’s not very well suited to what actually happens in school. This is shown both at the level of students in classrooms and at the level of the measurement instruments – the tests themselves. We now turn to what happens after a value-added score is generated to see the potential hazards of using this type of quantitative measure of teacher effectiveness as a tool in the educational workplace.

When New York City released its value added scores in 2012, the confidence intervals reported by the media – an average of 35 points for math and 54 points for English – showed exactly how unstable the rankings were be for a single year’s estimates. The largest margins of error (75 points for Math and 87 points for Reading) were found for teachers who were at or near the middle of the distribution. So while the estimate of a teacher’s relative effectiveness can be categorized as at, above, or below the average; the confidence interval could span all three categories (Santos, 2012). The inaccuracy of estimates is due to the small number of students each teacher works with even over the course of many years; even after five years of teaching in
a single grade, many teachers will have data for less than a hundred students. Further, as teachers often move in and out of grades and/or subjects, the amount of useful evaluative data will further decline.

Once scores are generated, they are typically used to rank teachers against their colleagues. This use of value added is problematic for the majority of teachers who are ranked because of problems with accuracy. While the ‘best’ and ‘worst’ teachers often rank consistently, most teachers, whose scores would place them in the middle of the distribution – for argument’s sake, within one standard deviation of the mean – are subject to often radical swings in their rankings (Corcoran, 2010; 2009; Amrein-Beardsley, 2008). In some cases, a fraction of a point’s difference in average value-added could push a teacher in the top third of the rankings (67th percentile) to the bottom third (33rd percentile) because both lie within one standard deviation. Depending on the district’s or state’s policy, that objectively minor shift could have major consequences for a teacher’s career (Winerip, 2012).

Beyond the composition, accuracy and propriety of value-added measures are concerns about the effects of the testing regime which it necessitates. Daniel Koretz (2000) goes into detail to explain how teachers can and do ‘game’ tests and what impacts this may have on education as a whole. The most blatant form of gaming is cheating, whereby teachers give students answers or change answers after the test is over. Widespread cheating in Atlanta, Georgia led one of that state’s leading newspapers to investigate cheating on a national level; indeed they found suspicious test score patterns – where improbable increases in one year were followed by drops in the following year – in school districts across the country (Vogell, Perry, Judd, & Pell, 2012). The authors of the investigative piece were very direct in noting that data driven accountability and ever-rising benchmarks mandated by No Child Left Behind figure prominently in the
increase of overt cheating on tests. Other specific cases of widespread cheating have been reported in New York, New Jersey, Philadelphia and Washington DC (Boyette, 2011; Otterman, 2011; Winerip, 2011b; 2011c).

Two more subtle forms of gaming are ‘coaching’ and ‘achievement transfer.’ Coaching in this context means teaching students tested concepts in the style that they appear on the exams; if division is tested horizontally – where the terms to be divided are listed left to right – it is specifically taught that way. Achievement transfer means spending more time on the subjects or topics that will be tested at the expense of non-tested subjects or, alternatively, focusing instruction on students who are likely to show larger score gains than on their peers who will not. While we could see these all happening with any standardized testing program, it would only be exacerbated if teachers felt that their job security was directly affected by the results of these exams.

A further problem concerns two types of institutional misattribution. The first is an obvious problem – many teachers teach subjects (or grades) where there is either no annual assessment or any at all. But in an accountability system where everyone must be evaluated on the basis of student test scores, this means that teachers may be evaluated based on score changes that have nothing to do with the subjects they teach. This was indeed the case in Tennessee, which was documented in the New York Times (Winerip, 2011). If these evaluations are consequential, as they are in Tennessee, teachers are liable to be promoted or rewarded, punished or fired, for score changes that have nothing to do with their own abilities.

---

17 Koretz found that when students who were taught to solve math problems vertically were subsequently given a test that presented them horizontally, scores dropped.
The second type misattribution can happen when cheating or other gaming happens. Because the scores in the prior year serve as the baseline for achievement growth in the following year, a teacher who games his students’ tests sets an artificially high benchmark for the next year’s teachers which is unlikely to be met. If the inflation of scores goes unnoticed, the teacher who is assigned to the previously gamed students will end up looking as if she added little or in fact detracted from student learning growth (Tuque 2012). It should be noted that this does not simply mean the prior year’s teacher needs to cheat; focused instruction on tested topics or matching instruction to test format could create the same artificially high pre-test scores. This type of misattribution should be cause for alarm given the recent string of cheating on high-stakes testing throughout the country.

It was noted above that value-added results in a relative assessment of teacher quality; this seems at face value to be an improvement over assessments based on ‘objective’ threshold, as these tend to privilege teachers with students who are closer to this benchmark at the outset. But while making the index relative to the student’s starting point is an improvement, the ultimate comparison is between teachers. In practice that means that some teachers will always be below average, even if what happens in the average classroom is ‘objectively’ good. Take the hypothetical example of ‘experienced’ teachers: when compared to newer teachers, they tend to perform better; however, when compared to one another, some if not many experienced teachers who are generally good may fall below average on this type of assessment. If Value-Added is used to determine pay increases, bonuses, and discipline, the unintended consequence could be to chase good teachers from the grades and subjects where their efforts are most valuable. Hypothetically, you could have quality experienced math teachers opting for Kindergarten classrooms for fear of assessment.
Sociological trouble with Value-Added: An Authority Paradox

A final potential pitfall that has not been often mentioned is a paradox of classroom authority. This returns us to our starting point – work measurement. Stake (1971) mentions this potential in an article written about performance contracting in education. He makes the plausible argument that in a situation where teachers’ evaluations are dependent on student test outcomes, students have an unusual amount of power. If students can achieve a wide variety of scores on assessments when giving their best efforts, it is certainly possible for them to achieve an even wider variety of results without such. Simply put and to quote Stake, “bad performances are in their repertoire.” While this is an unlikely situation, it is a consequence that needs to be considered, especially as the consequences of these tests on teachers become widely known. To move to a factory analogy, Burawoy (1979) points out those personnel not subject to piece-rates were often very consequential for the pace of piece-rate work. Such calculations of efficiency, piece-rates and value-added, do not take into account this problem of inter-subjectivity; the output is treated as if it were the sole responsibility of the individual worker. In the case of the teacher, the very object of their action is itself a self-directed actor, which raises this problem from an occasional incident to a systemic flaw.

But I want to take the concept of an authority paradox back to the classroom using a largely-forgotten sociological work on teaching from Willard Waller.¹⁸ I will conclude by considering Waller’s notion of classroom authority, its relation to Durkheim’s conception of authority and education, and the impact of high-stakes testing on this relationship.

¹⁸ I first discovered Waller’s studies of teaching while reading David Labaree’s (2010) work on school reform movements.
To this point, we’ve shown that the assumptions of value-added assessment betray many important classroom realities. Indeed, this mismatch between model assumptions and actual school and classroom life are at the heart of the American Statistical Association’s and American Education Research Association’s official objections to this sort of testing-driven assessment (American Statistical Association, 2014; American Educational Research Association, 2015). But the effects of VAA on classroom authority and the interaction between students and teachers are not as well-attended. Willard Waller came to sociology as a former high school teacher, and closely examined the social and psychological situation of teachers and students in the classroom.

For Waller, schooling is a distinctive social situation because of its mode of authority. School is organized around ‘despotic authority,’ in which teachers and school administrators unilaterally dominate students. But he characterizes that authority as in a state of ‘perilous equilibrium.’ While power in the school is theoretically vested in administration, and practically vested in teachers, “The control of school affairs…rest[s] upon the consent, mostly silent, of the governed.” (Waller, 1932/2014, p. 12) Thus, Waller places the ultimate power in schools with students, “the most tractable but most unstable members of the [school] community.” (p. 10) Furthermore, parents and the wider community place constraints on teachers’ and administrators’ authority in the classroom, compounding the fragile classroom situation. Thus, one main task question in his Sociology of Teaching is “How can the teacher control school life at all?” (p. 3)

Waller discusses various factors which shape classroom authority. School imposes an artificial social situation with pre-determined goals. Students are compelled to attend school and have no say in its content or conduct. As such, schooling is organized around ‘institutionalized leadership.’ The teacher holds authority because of herinstitutional position. This is distinct
from personal or charismatic leadership, in which leaders emerge from among groups by their actions. Because of the institutionalized nature of her position, the teacher is obliged to maintain social distance from her students through a system of formal relations. This formality keeps the individual personalities of teachers and students from coming into contact. Waller sees the teachers’ use of personal authority as a disciplinary risk, as it would reduce social distance and thereby degrade institutional authority – which he regards as far less compelling than personal authority. Social distance also keeps students from investing too much emotional energy into the classroom situation.

But why keep students from investing energy in their schooling? To keep the inherent antagonism of the classroom under control. Beneath the formal classroom relationship lays a fundamental hostility between the ordered world of adults and the spontaneous desires of children. Waller explains this hostility through Simmel’s conception of the feud, in which antagonisms are based on group status rather than personal motives. Teachers are charged with pushing students into the world of adults, which affronts the world constructed by groups of children. The despotic arrangement of schools amplifies this enmity by preventing teachers and students from engaging interpersonally – denying students the sort of personal connection they desire. Even as they adhere to the formal rules, students express hostility. Although school rules favor the teacher, students resist by mechanically obeying rules or engaging in disobedience not covered by the rules. Students thereby hollow out the meaning and intention of the rules. Waller suggests that this subversion of classroom rules can only be minimized if teachers command

---

19 Waller is not in favor of this structure of schooling. Institutional leadership tends toward formalism, where the institution serves its own narrow ends without regard to its intended social function.

20 Simmel develops this concept in *Conflict and the Web of Group-Affiliations* (1922/1955).
some amount of personal influence over students beyond the institutional power they are assigned.

This general condition of institutional leadership is troubled by another feature of schooling. A classroom full of students constitutes a crowd. This group setting makes order more unstable, and obliges the teacher to hold the audience, keeping her relationship with its members “significant to the exclusion of others.” (p. 160) The classroom crowd can transform into a mob if sufficiently incited. When this happens, inhibitions are lowered and the fundamental hostility between students and teacher is more likely to reach the surface as students react to one another rather than to the teacher. This transformation must be actively suppressed. The teacher seeks to maintain students’ crowd orientation by avoiding emotional reactions to students, maintaining seamless classroom routines, and isolating individuals who might incite a mob feeling. Losing control of a class is one of the most traumatic events in the teacher’s career, and it either reinforces her commitment to classroom authority or destroys her career altogether.

Thus personal engagement is both necessary and dangerous in the classroom – teachers cannot afford to breach formality for the sake of discipline, but discipline without personal influence breeds discontent. Thus successful teachers must “filter personal influence through the sieve of the institutional relationship.” (p. 190) The successful teacher carefully utilizes formal rules and sanctions, and only infrequently (in crisis situations) transcends the formal setting to engage in personality contact with students. These rare instances may take the form of emotional responses like anger or of frank appeals to individual students.

---
21 The theory of crowd psychology [e.g., LeBon (1895/1969); Tarde (1898/1969)] had an important impact on the early development of Sociology.
Waller’s emphasis on the student-teacher relationship can be traced further back to Emile Durkheim’s sociology of education. In *Moral Education*, Durkheim’s most comprehensive statement on schooling, he defines education as a project of secular morality. He argues that the family and the church cannot, as they once had, be responsible for transmitting morals befitting modern society. Durkheim believes that “we must discover the rational substitutes for those religious notions that for a long time have served as a vehicle for the most essential moral ideas.” (Durkheim, 1961, p. 9) But as he posits the necessity of rational notions to replace religious ones, Durkheim is careful not to push this emphasis so far as to eliminate the sacred character of the educational process.

“If then, in rationalizing education, we do not retain this [sacred] character and make it clear to the child in a rational manner, we will only transmit to him a morality fallen from its natural dignity. At the same time, we will risk drying up the source from which the schoolmaster himself drew a part of his authority and also a part of the warmth necessary to stir the heart and stimulate the mind.” (p. 13)

For Durkheim, authority is strictly relational, bound up in the interactions among individuals in such a way that some feel others to be superior. Authority figures are so because they embody (or are seen to embody) the values of society. As for the authority of the teacher, Durkheim feels this depends on the teacher’s respect for her own role as an educator which radiates out to the students. This veneration of the role is distinct from a sense of personal superiority, which Durkheim (like Waller) sees as detrimental to the educational process. 22

22 Durkheim and Waller differ on why personal authority is dangerous. Durkheim sees school discipline as an educational device in itself. This is based on his contention that education is a process of moral formation, in which children learn to conduct themselves for the benefit of society. For this process to be successful, the authority of the teacher has to appear to students as impersonal, a stand-in for the social orientation that students must internalize. Waller agrees nominally with the importance of morality for the school teacher, but sees personal authority as dangerous primarily to the orderly functioning of the classroom.
“…the dominant part played by the teacher in generating this feeling [of respect for the rules] – the personal role that he plays – entails a danger we must guard against…lest the child develop the habit of associating the rule itself too narrowly with the person of the teacher…” (p. 156)

Thus Waller’s focus on the educator’s need for personal authority is located firmly in the sociological tradition. Returning to the broader question, the classroom situation is as perilous as Waller describes; his references to feuds, deep-seeded hostility, crowds and mobs are all understandable to the experienced teacher. Now consider an invasive teacher evaluation system which depends upon student performance from the vantage point of the student-teacher interaction. It would not take long for students to understand their role in these assessments: their performance on examinations determines whether their teacher is considered effective or not. 23 Teachers could disclose this actively by suggesting that students ‘not try hard’ on pre-tests to improve growth scores. Even without overt declarations, increased emphasis on test preparation could further indicate a teacher’s personal stake in narrow student outcomes. For the vindictive student, this could appear as an opportunity to get back at a teacher for real or perceived slights. Even among compliant students, the sense of who controls the classroom could be altered. In short, teacher evaluations so closely tied to students’ outcomes requires much more student cooperation in the classroom. This impacts more than just the potential for disruptions, as serious as that can be for teachers and students alike; it undermines the notion of authority that characterizes the classroom. Under such an evaluation regime, students may not perceive the teacher as vested with the authority of the institution, as she is as much a subject of external domination as they are. This leveling of teachers and students – now equal before the eyes of the

23 Rothstein (Rothstein J., 2011) mentions such hazards of measurement in his review of an early stage report of the Gates Foundation Measures of Effective Teaching study. “But the MET study’s inability to examine how teachers, students and administrators respond to the use of MET measures for high stakes decisions limits what the MET project can tell us about the utility of the measures for real world use.” (p. 7)
exam – could further jeopardize the already perilous classroom situation. Whether or not one agrees with the goals and methods of traditional schooling, it is clear that such invasive evaluation systems problematizes that system.

But beyond the authority question, there is a further problem which is endemic to such evaluation which Waller would recognize. Value-Added assessment requires a lot of testing. As noted above, proponents argue that there should be pre- and post-testing each year in every subject to accurately capture teacher effects. While he far predates the large expansion of testing documented in Chapter 3, Waller addressed testing and grading as they relate to classroom authority and the broader project of education:

“The by introducing an element of the artificial and the forced, by placing the emphasis on upon the false, objective symbol of the subjective reality of learning, it vitiates the relationship of teacher and taught.” (p. 372)

The relationship between teacher and taught Waller describes is challenging, but it is also vital to the enterprise of education. Testing harms that relationship by mistaking a representation for the thing itself. Curiously, this anticipates Donald Campbell’s (1975) law of social indicators by half a century, but in a much more important way. Over time, excessive evaluation saps students’ desire for education. Waller says that while the primary school student desires the teacher’s personal influence (which the primary school teacher avoids), by the time students arrive in college – where teachers seek to exert personal influence on young adults – the students have become institutionalized and desire the narrow formalism of the grading system (p. 368). This troubling paradox is only exacerbated as the scope and importance of testing grow. Whereas teachers in the past may have been able to use testing at their discretion and thus tread delicately
over this paradox, teachers under a high-stakes testing regime cannot afford to deviate from the script of preparation and evaluation.\textsuperscript{24}

So far, we’ve considered the theoretical model of value-added against the practical realities of the schools and classrooms in which it would be applied. By way of early theories in the sociology of education, we have suggested some short- and long-term negative consequences such an evaluation regime would have. We now turn to consider empirical Value-Added data in order to demonstrate some of the problems outlined here. If VA has troubling theoretical characteristics, and some are already empirical demonstrable, we will then move to consider how and why it has become a preferred mode of educational evaluation.

\textsuperscript{24} In chapter 3, we will see that even some of the most ardent standardized testers were careful to frame their products as something that should be tailored to fit by professional educators.
Tests are, by definition, indirect and incomplete indicators of students’ knowledge of subject matter. They do not represent the quality or quantity of education received in the schooling process, and certainly do not measure anything about that process or the individual responsible for it. Value-Added makes inferences based on a third-order abstraction, bypassing the complexities that bring us from classroom to test score.

Figure 1: Value-Added and the teacher-student-test score chain
Chapter 3: Value-Added by the Numbers: A Quantitative Presentation Using NYC Teacher Evaluation Data

Having discussed the problems with VAA in theory, I move to demonstrating some of the major empirical problems with Value Added assessment of individual teachers using existing data, as well as adding some potential problems not considered by others.

There are three issues demonstrated below. First, the concept of temporal instability noted by Darling-Hammond et al. (2012) and Sean Corcoran (Corcoran, 2010), among others – that teacher Value-Added rankings vary over time well beyond what one would expect of a reliable rating system. Secondly, making use of teachers who had two records or more in a given year, I show that value-added is not only unstable temporally, but also that it is highly variable for an individual who teaches more than one class in a given year. Finally, I demonstrate the problem of “persistent school effects;” despite the fact that New York City value-added measures control statistically for school level variables in the estimation of teacher value-added, I find that school characteristics are still highly predictive of a teacher’s value-added ranking. The cumulative evidence here suggests that teacher ranking according to value-added scores is both imprecise in assuming that abilities are relatively stable and improper given that characteristics of schools are beyond teachers’ capacity to change. This is problematic given that proponents of VA suggest that among existing evaluation methods, VA is unique in holding teachers accountable only for things they can control (Harris D. N., Value-Added Measures in Education: What Every Educator Needs to Know, 2011).

25 Also see Goldstein (1997) and McCaffrey et al. (2008).
26 The problem of school-level sorting is also noted by Rothstein (2009; 2008).
Data Source: NYC Value-Added FOIL Data

Under Mayor Michael Bloomberg and schools Chancellor Joel Klein, New York City launched a pilot program in the use of value-added assessment for teacher evaluation. The program was conducted by the Value-Added Research Center at the University of Wisconsin-Madison and cost an estimated $3.6 million (Otterman, 2010). New York State has since contracted the American Institutes of Research Value-Added division to construct its teacher evaluation system.

The data come from a Freedom of Information Law (FOIL) request made by *The New York Times* and a dozen other New York City media outlets to the NYC Department of Education in September of 2010 (Santos, 2012). The United Federation of Teachers (UFT) sued to keep the named teacher rankings private, and the subsequent litigation delayed the release until February of 2012. In September of 2011, New York City announced that it would no longer personally oversee the creation of teacher data reports which included these value-added rankings; that function has now passed to the state level (Phillips A. , 2012).

The data were made available for download by television station New York 1, which was party to the FOIL request to the Department of Education.27 There were three data files (and three supplemental files) corresponding to three years 2009-10, 2008-09, and 2007-08. Each of those files contains Value Added measures and rankings based on three or four years of data (called multi-year) a measure and ranking based on individual year data (called single-year). The first file also contains single-year estimates for 2005-06 and 2006-07. Thus, I have five single-year and three multi-year data points. Math and English Language Arts value-added scores are

---

calculated for NYC teachers of grades 4 through 8; each grade/subject combination generates its own score, so most teachers in fact receive more than one score per year. Any data point based on fewer than 10 students’ scores is omitted from publicly-available data because of the Family Educational Rights and Privacy Act (FERPA) of 1974.

The table below presents the number of cases available for each data point across the two tested subjects. Each of the analyses presented below will make use of different combinations of data points. The first three single-year estimates come from a single file provided in the FOIL request, and are thus used for the analyses of temporal stability and of persisting school effects.\(^{28}\) The analysis of intra-teacher stability makes use of the three multi-year data points in both subjects.

Table 1: Number of Records for Data Points in the FOIL data file

<table>
<thead>
<tr>
<th>Data Point</th>
<th>N of Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005-06 single year</td>
<td>6,488</td>
</tr>
<tr>
<td>2006-07 single year</td>
<td>9,022</td>
</tr>
<tr>
<td>2007-08 single year</td>
<td>16,707</td>
</tr>
<tr>
<td>2008-09 single year</td>
<td>16,044</td>
</tr>
<tr>
<td>2009-10 single year</td>
<td>15,630</td>
</tr>
<tr>
<td>2005-07 multi-year</td>
<td>16,707</td>
</tr>
<tr>
<td>2006-08 multi-year</td>
<td>10,779</td>
</tr>
<tr>
<td>2007-09 multi-year</td>
<td>11,289</td>
</tr>
</tbody>
</table>

The FOIL data were merged with school-level characteristics drawn from parallel years of the New York City School Progress Reports; both data files used the same unique identifier for schools. These reports contained variables including school enrollment size and attendance rate, the percentage of Title I students (a proxy measure of economic disadvantage), the percentage of students who were black or Hispanic, and an environment score which incorporates student,\

\(^{28}\) New York 1 did not provide unique teacher identification variables, which made it difficult to merge the three separate FOIL data files.
staff, and parent perceptions of school safety and atmosphere.\textsuperscript{29} Below, I show the number and proportion of cases in each subject/year of the FOIL data that were successfully merged with Progress Report data.

Table 2: Merge Results for FOIL and NYC School Progress Report data

<table>
<thead>
<tr>
<th>VA Data Year/Subject</th>
<th>N of Cases</th>
<th>% matched (N matched)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006 ELA</td>
<td>8,779</td>
<td>98% (8,650)</td>
</tr>
<tr>
<td>2006 MATH</td>
<td>8,660</td>
<td>98% (8,526)</td>
</tr>
<tr>
<td>2007 ELA</td>
<td>8,547</td>
<td>99% (8,474)</td>
</tr>
<tr>
<td>2007 MATH</td>
<td>9,404</td>
<td>99% (9,333)</td>
</tr>
<tr>
<td>2008 ELA</td>
<td>8,580</td>
<td>99% (8,498)</td>
</tr>
<tr>
<td>2008 MATH</td>
<td>9,409</td>
<td>99% (9,332)</td>
</tr>
</tbody>
</table>

To demonstrate the temporal instability of value-added rankings, I use two techniques. First, I calculate changes in rankings between single-year estimates and produce histograms of absolute changes. This is used to discern the extent to which teachers’ rankings bounce over time. Second, I use regression models to investigate what variables are related to the bouncing of teacher rankings.

To show the persistence of school effects in determining teacher value-added, I first utilize Analysis of Variance to show that there is a significant amount variance in value-added rankings between schools. Then, using a statistical tool known as a sheaf coefficient, I demonstrate that school characteristics when taken together are a powerful predictor of changes in teacher value-added. This is a significant finding given that the rankings themselves are supposed to control for school characteristics.

\textsuperscript{29} Documentation on the NYC school progress reports can be accessed at: http://schools.nyc.gov/Accountability/tools/report/default.htm#Methods
Finally, I look at within-teacher, within-subject variability at a single point in time. To do so, I first plot a teacher’s two rankings and report correlation coefficients. Then, I use *kappa*, a measure of inter-coder agreement to assess whether value-added rankings are any more reliable at evaluating teachers than random chance.

Findings

**VA “Bouncing:” Temporal Stability of Value-Added Rankings**

The performance categories used here are the same defined by the methodology report produced by the Value Added Research Center (VARC) which was contracted by the New York City Department of Education to collect and analyze these data. ‘Low’ corresponds to rankings below the 5th percentile; ‘Below Average’ to rankings between the 5th and 24th percentile; ‘Average’ to rankings between the 25th and 74th percentile; ‘Above Average’ to rankings between the 75th and 94th percentile; and high refers to rankings at the 95th percentile and above.30

Tables 3 through 8 below show teachers’ movement among performance categories between three single-year data points. I refer to this phenomenon as “bouncing.” For simplicity, only teachers with all three years of data were included in these tabulations.

---

30 In many policy-oriented papers on VA, the top and bottom five percentiles are singled out as “highly effective” or “grossly ineffective.” The latter category is the focus of many arguments for teacher dismissal policies, while the former is often discussed in terms of merit pay bonuses (Chetty, Friedman, & Rockoff, 2011a; Hanushek E., 2009).
Table 3: Ranking movement between year 1 and year 2 – All Math Instructors (N=2,985)

<table>
<thead>
<tr>
<th>05-06 category</th>
<th>Low</th>
<th>Below Average</th>
<th>06-07 category (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Average</td>
</tr>
<tr>
<td>Low</td>
<td>0</td>
<td>42.9</td>
<td>57.1</td>
</tr>
<tr>
<td>Below Average</td>
<td>0.7</td>
<td>28.8</td>
<td>67.9</td>
</tr>
<tr>
<td>Average</td>
<td>0.4</td>
<td>11.2</td>
<td>76.0</td>
</tr>
<tr>
<td>Above Average</td>
<td>0.3</td>
<td>5.2</td>
<td>56.4</td>
</tr>
<tr>
<td>High</td>
<td>0</td>
<td>2.0</td>
<td>46.0</td>
</tr>
</tbody>
</table>

The first observation from table 3 is that the majority of teachers do not move out of the ‘average’ performance category between any two years. Indeed, this is the case across all six of these tables. The width of the ‘Average’ category (50 percentile points) makes this unsurprising.

What we observe is that even if a teacher was classified in the ‘High’ or ‘Low’ categories in the first year, she is very likely to be classified as Average in the following year.

Table 4: Ranking Movement between year 2 and year 3 – All Math Instructors (N=2,985)

<table>
<thead>
<tr>
<th>06-07 category</th>
<th>Low</th>
<th>Below Average</th>
<th>07-08 Category (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Average</td>
</tr>
<tr>
<td>Low</td>
<td>0</td>
<td>30.8</td>
<td>69.2</td>
</tr>
<tr>
<td>Below Average</td>
<td>2.0</td>
<td>30.6</td>
<td>62.6</td>
</tr>
<tr>
<td>Average</td>
<td>0.4</td>
<td>12.0</td>
<td>74.9</td>
</tr>
<tr>
<td>Above Average</td>
<td>0.5</td>
<td>3.4</td>
<td>58.2</td>
</tr>
<tr>
<td>High</td>
<td>0.0</td>
<td>5.1</td>
<td>37.3</td>
</tr>
</tbody>
</table>

The same pattern is found in table 4. No math teacher who was ranked ‘Low’ in the 2006-07 school year was found in the same category in the following year. The majority of ‘Below Average’ and ‘Above Average’ teachers in 2006 drifted into the middle category in 2007.

Teachers ranked in the ‘High’ category were unlikely to fall below average ranking, but over a third fell into the ‘average’ category.
Examining movement among math teachers over a two-year period from 2005-06 to 2007-08 (Table 5), we again see the same general trends. Most math teachers are ranked ‘Average’ at one or the other time point, and even those ranked ‘High’ or ‘Low’ more often than not return to the broad middle.

Table 5: Ranking Movement between year 1 and year 3 – All Math Instructors (N=2,985)

<table>
<thead>
<tr>
<th>05-06 category</th>
<th>Low</th>
<th>Below Average</th>
<th>07-08 Category (%)</th>
<th>Average</th>
<th>Above Average</th>
<th>High</th>
<th>Totals (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0</td>
<td>42.9</td>
<td>42.9</td>
<td>14.3</td>
<td>0</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Below Average</td>
<td>1.9</td>
<td>26.6</td>
<td>66.4</td>
<td>5.2</td>
<td>0</td>
<td>271</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>0.5</td>
<td>12.9</td>
<td>73.7</td>
<td>11.9</td>
<td>1.1</td>
<td>2,251</td>
<td></td>
</tr>
<tr>
<td>Above Average</td>
<td>0</td>
<td>5.4</td>
<td>57.9</td>
<td>30.5</td>
<td>6.2</td>
<td>406</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>0</td>
<td>0</td>
<td>50.0</td>
<td>30.0</td>
<td>20.0</td>
<td>50</td>
<td></td>
</tr>
</tbody>
</table>

With English Language Arts (ELA), the drift to the middle is again the dominant trend, though we see a bit more extreme movement for those ranked in the top categories. In table six, we see that eight percent of ‘Below Average’ teachers in 2005 moved into the top two categories in the following year; likewise, about six percent of those ranked ‘above average’ moved to the bottom two categories in 2006.

Table 6: Ranking movement between year 1 and year 2 – All ELA (N=2,795)

<table>
<thead>
<tr>
<th>05-06 Category</th>
<th>Low</th>
<th>Below Average</th>
<th>06-07 Category</th>
<th>Average</th>
<th>Above Average</th>
<th>High</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0</td>
<td>7.7</td>
<td>61.5</td>
<td>30.8</td>
<td>0</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>Below Average</td>
<td>0.4</td>
<td>11.6</td>
<td>80</td>
<td>7.5</td>
<td>0.4</td>
<td>225</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>0.1</td>
<td>6.6</td>
<td>87.1</td>
<td>5.8</td>
<td>0.5</td>
<td>2,324</td>
<td></td>
</tr>
<tr>
<td>Above Average</td>
<td>0.3</td>
<td>6.0</td>
<td>82.9</td>
<td>10.7</td>
<td>0</td>
<td>298</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>0</td>
<td>0</td>
<td>72.0</td>
<td>20.0</td>
<td>8.0</td>
<td>25</td>
<td></td>
</tr>
</tbody>
</table>
Table 7: Ranking Movement between year 2 and year 3 – All ELA (N=2,795)

<table>
<thead>
<tr>
<th>06-07 Category</th>
<th>Low</th>
<th>Below Average</th>
<th>Average</th>
<th>Above Average</th>
<th>High</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Below Average</td>
<td>0</td>
<td>17.7</td>
<td>74.0</td>
<td>8.3</td>
<td>0</td>
<td>192</td>
</tr>
<tr>
<td>Average</td>
<td>0.7</td>
<td>9.5</td>
<td>77.3</td>
<td>11.6</td>
<td>0.9</td>
<td>2,399</td>
</tr>
<tr>
<td>Above Average</td>
<td>1.6</td>
<td>8.0</td>
<td>64.2</td>
<td>23.0</td>
<td>3.2</td>
<td>187</td>
</tr>
<tr>
<td>High</td>
<td>0</td>
<td>7.1</td>
<td>35.7</td>
<td>28.6</td>
<td>28.6</td>
<td>14</td>
</tr>
</tbody>
</table>

The same pattern is observed between 2006-07 and 2007-08, shown in table 7. The great majority of teachers were ranked in the ‘Average’ category in at least one of the two years. A small but not inconsiderable proportion of teachers ranked ‘Above Average’ or ‘Below Average’ moved to the opposite side of the distribution. When we examine movement among ELA teachers over the two-year period between 2005 and 2007 (Table 8), we see the same patterns as before.

Table 8: Ranking Movement between year 1 and year 3 – All ELA (N=2,795)

<table>
<thead>
<tr>
<th>05-06 Category</th>
<th>Low</th>
<th>Below Average</th>
<th>Average</th>
<th>Above Average</th>
<th>High</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>7.7</td>
<td>0</td>
<td>69.2</td>
<td>15.4</td>
<td>7.7</td>
<td>13</td>
</tr>
<tr>
<td>Below Average</td>
<td>0.4</td>
<td>13.3</td>
<td>77.3</td>
<td>8.0</td>
<td>0.9</td>
<td>225</td>
</tr>
<tr>
<td>Average</td>
<td>0.7</td>
<td>10.5</td>
<td>77.4</td>
<td>10.8</td>
<td>0.6</td>
<td>2,324</td>
</tr>
<tr>
<td>Above Average</td>
<td>0.7</td>
<td>4.4</td>
<td>67.8</td>
<td>23.8</td>
<td>3.4</td>
<td>298</td>
</tr>
<tr>
<td>High</td>
<td>0.0</td>
<td>4.0</td>
<td>40.0</td>
<td>40.0</td>
<td>16.0</td>
<td>25</td>
</tr>
</tbody>
</table>

To provide a more nuanced picture of year-to-year movement, I plot the absolute change in teacher rankings between years. This allows us to see changes not just between categories, but the magnitude of the swings in ranking from year to year. The first set of graphs refers to changes between 2005-2006 and 2006-2007.
Across all four frequency charts in Figure 2, we see that a large proportion of teachers had modest ranking changes of 10 percentage points or less. But a fair proportion had high absolute changes that were 30 points or more. Many of these changes would not often have had consequences for teachers had the rankings been used for evaluation purposes. But from the perspective of a measurement system, these often dramatic ranking changes suggest that there is
not a stable characteristic underlying these assessments.\textsuperscript{31} If there were, we would expect a much steeper decline from left to right in the chart, with very few teachers experiencing high absolute changes in ranking.

Figure 3: Absolute ranking change between Year 2 and Year 3

Figure 3 presents the same array of four histograms, but for changes between 2006 and 2007. Particularly in Middle School ELA and Mathematics, if we set 20 percentage points as a threshold, nearly equal proportions of the data are on either side. The very shallow slope of the

\textsuperscript{31} From a working conditions perspective, the stability of a ranking system can also impact teacher morale. If teachers perceive assessments as having little direct relationship to the work they do, it can lead to feelings of hopelessness. Rothstein (2012) notes this in his simulation study of teacher quality policy.
bottom two panels suggests that the reports of wide confidence intervals in a single teacher’s data report bear out in subsequent years.

Figure 4: Absolute ranking change between Year 1 and Year 3

Elementary ELA

Elementary Mathematics

Middle School ELA

Middle School Mathematics

Figure 4 again presents the same frequency histograms, but for changes over a 2-year period between 2005 and 2007. Here again we see very shallow slopes in most of the histograms which suggest the prevalence of very drastic absolute changes in individual teachers’ rankings in this two-year period.
Value-Added “Swaying:” Within-Teacher Variation in VA Rankings

The previous demonstrations focused on year-to-year changes in teacher rankings – bouncing – and showed them in some cases to be quite dramatic, even if very few would have had consequences for teachers in terms of either merit pay or dismissal in a ‘live’ assessment system. Now we turn to a different sort of variation in rankings – within teacher variation – which I call “swaying.” This analysis takes teachers who received more than one ranking in a single year, in a single subject at the same school. Most of these therefore come from middle school teachers who taught multiple grades in the same year. For simplification purposes, I took only teachers who received exactly two value-added rankings and excluded those who received three or more. The purpose of the following two sets of analysis is to demonstrate that even within one year, a teacher’s value-added ranking is highly unstable. Here, instead of using one-year data points, we use the three-year data points which are recommended for use in teacher evaluation by VARC.³²

³² Q: Which teachers received Teacher Data Reports?
A: Reports were generated for approximately 18,000 English language arts and math teachers who taught 4th-8th grade students in 2007-08, 2008-09, and 2009-10. Teachers did not receive these reports if they did not teach a subject that culminated in a state math or English exam, taught for less than the full year, or did not teach a minimum number of students required to receive a report. Teacher Data Reports were calculated for each grade and subject individually. One teacher may therefore have multiple reports (e.g., a teacher who teaches 4th grade math and English).
Figure 5: Scatterplot of ELA teachers’ rankings in one year in two separate classrooms

Figure 5 presents a scatterplot showing a teacher’s pair of rankings in a single year. Given the same time and same subject, we would expect to see at least some correlation in their ranking – but the correlation coefficient, Pearson’s r, is a non-significant .08. This means that there is very weak non-significant relationship between how a teacher is ranked based on her ‘performance’ in one classroom and her performance in another classroom in the exact same time period. We see data points in both the bottom right and top left corners, representing teachers who were among the ‘best’ based on their performance in one grade level and among the ‘worst’ based on their performance in another. This is hard to believe given that it is the same teacher at the same point in time.

Pearson’s r statistic = .0796, p=.1595
The data for dual-grade math instructors paints a similar picture. Notably, there is a statistically significant Pearson’s r of .350 (p<.001); but this figure represents a weak correlation between the two variables. The diagram again shows a wide distribution, including values in the top-left and bottom-right corners where correlated data should not appear. This suggests that the relationship between a teacher’s performance rankings in one grade and her performance rank in another grade in the same year at the same school are at best weakly correlated.

As another way of examining teachers with multiple value-added rankings in a single year, we can use a statistic of inter-observer agreement. In medical research, these sorts of measures are used to determine the extent to which two observers agree on the interpretation of various patient records; but they can theoretically be used “in any situation in which two or more independent observers are evaluating the same thing” (Viera & Garrett, 2005). The statistic that measures the rate of agreement is called kappa.
Table 9: ELA teachers’ within-subject, within-year rankings with kappa statistic

<table>
<thead>
<tr>
<th>Classroom 1 /Classroom 2</th>
<th>Low</th>
<th>Below Average</th>
<th>Average</th>
<th>Above Average</th>
<th>High</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Below Average</td>
<td>0</td>
<td>9</td>
<td>38</td>
<td>8</td>
<td>1</td>
<td>49</td>
</tr>
<tr>
<td>Average</td>
<td>0</td>
<td>37</td>
<td>146</td>
<td>27</td>
<td>5</td>
<td>215</td>
</tr>
<tr>
<td>Above Average</td>
<td>0</td>
<td>4</td>
<td>29</td>
<td>3</td>
<td>3</td>
<td>39</td>
</tr>
<tr>
<td>High</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 9 shows that 159/314 cases (51 percent) agree across classrooms. We would expect 160 cases to agree by chance alone. This generates a kappa value of -.0065 (p=.56). Thus if we take two rankings of the same teacher in the same year in English, the value-added rankings correlate no better than if they had been assigned by chance.

Table 10: Math teachers’ within-subject, within-year rankings with kappa statistic

<table>
<thead>
<tr>
<th>Classroom 1 /Classroom 2</th>
<th>Low</th>
<th>Below Average</th>
<th>Average</th>
<th>Above Average</th>
<th>High</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Below Average</td>
<td>2</td>
<td>19</td>
<td>35</td>
<td>5</td>
<td>3</td>
<td>64</td>
</tr>
<tr>
<td>Average</td>
<td>5</td>
<td>49</td>
<td>101</td>
<td>32</td>
<td>13</td>
<td>200</td>
</tr>
<tr>
<td>Above Average</td>
<td>2</td>
<td>4</td>
<td>29</td>
<td>16</td>
<td>6</td>
<td>57</td>
</tr>
<tr>
<td>High</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>8</td>
<td>0</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 10 shows that 137/346 cases (40%) agree across classrooms. We would expect 127 cases to agree by chance alone. This generates a kappa value of .0450 (p=.09). Thus if we take two rankings of the same teacher in the same year in mathematics, the value-added scores correlate no better than if they had been assigned by chance.

Schools Matter – The Persistence of School-level variables in VA rankings

The third set of findings concerns the charge that school characteristics continue to affect value-added scores even though the scores themselves are meant to account for school-level variables
in the models. We begin with ANOVA tests to get baseline estimates of whether or not there is significant variation in value added ranking changes between schools.

Table 11: ANOVA tests on change in value added ranking between 2005 and 2006 with school as the grouping variable

<table>
<thead>
<tr>
<th>Subject</th>
<th>R²</th>
<th>SST</th>
<th>SSB</th>
<th>SSW</th>
<th>F</th>
<th>Sig.</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elementary Math</td>
<td>.419</td>
<td>1070027.8</td>
<td>448096.2</td>
<td>621931.6</td>
<td>1.63</td>
<td>&lt;.001</td>
<td>1,999</td>
</tr>
<tr>
<td>Elementary ELA</td>
<td>.474</td>
<td>997759.1</td>
<td>463056.8</td>
<td>514702.2</td>
<td>1.90</td>
<td>&lt;.001</td>
<td>1,883</td>
</tr>
<tr>
<td>Middle School Math</td>
<td>.359</td>
<td>483900.2</td>
<td>173552.9</td>
<td>310347.3</td>
<td>1.52</td>
<td>&lt;.001</td>
<td>986</td>
</tr>
<tr>
<td>Middle School ELA</td>
<td>.391</td>
<td>625158.7</td>
<td>244639.6</td>
<td>380519.1</td>
<td>1.79</td>
<td>&lt;.001</td>
<td>912</td>
</tr>
</tbody>
</table>

Tables 11 through 13 report the results of a series of ANOVA tests. The groups are individual schools and so the hypothesis tested is: do changes in value added ranking vary by school? For table 11, in each subject and school level, we have a statistically significant F-value of between 1.5 and 1.9. We can interpret this as suggesting that changes in value added ranking between the 2005 and 2006 school years did indeed vary by school. In all cases the Sum of Squares Within (SSW) was greater than the Sum of Squares Between (SSB), which suggests that there was more variation in value-added ranking changes within an individual school than between schools.

Table 12: ANOVA tests on change in value added ranking between 2006 and 2007 with school as the grouping variable

<table>
<thead>
<tr>
<th>Subject</th>
<th>R²</th>
<th>SST</th>
<th>SSB</th>
<th>SSW</th>
<th>F</th>
<th>Sig.</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elementary Math</td>
<td>.409</td>
<td>1033390.7</td>
<td>422667.4</td>
<td>610723.3</td>
<td>1.57</td>
<td>&lt;.001</td>
<td>1,999</td>
</tr>
<tr>
<td>Elementary ELA</td>
<td>.427</td>
<td>1018344.2</td>
<td>434627.3</td>
<td>583717.2</td>
<td>1.58</td>
<td>&lt;.001</td>
<td>1,883</td>
</tr>
<tr>
<td>Middle School Math</td>
<td>.359</td>
<td>572425.5</td>
<td>205717.1</td>
<td>366708.5</td>
<td>1.52</td>
<td>&lt;.001</td>
<td>986</td>
</tr>
<tr>
<td>Middle School ELA</td>
<td>.368</td>
<td>602509.5</td>
<td>221398.7</td>
<td>381110.8</td>
<td>1.62</td>
<td>&lt;.001</td>
<td>912</td>
</tr>
</tbody>
</table>

Table 12 again shows all statistically significant F statistics, which suggest that changes in value added ranking between the 2006 and 2007 school years did indeed vary significantly between
schools. Again SSW values were greater than SSB values, which suggest that there was more variation in value-added ranking changes within individual schools.

Table 13: ANOVA tests on absolute change in value added ranking between 2005 and 2007 with school as the grouping variable

<table>
<thead>
<tr>
<th></th>
<th>R²</th>
<th>SST</th>
<th>SSB</th>
<th>SSW</th>
<th>F</th>
<th>Sig.</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elementary Math</td>
<td>.415</td>
<td>1096626.8</td>
<td>455066.9</td>
<td>641559.9</td>
<td>1.61</td>
<td>&lt;.001</td>
<td>1999</td>
</tr>
<tr>
<td>Elementary ELA</td>
<td>.4299</td>
<td>1045034.1</td>
<td>595762.3</td>
<td>449271.8</td>
<td>1.60</td>
<td>&lt;.001</td>
<td>1883</td>
</tr>
<tr>
<td>Middle School Math</td>
<td>.389</td>
<td>583267.03</td>
<td>227036.5</td>
<td>356230.5</td>
<td>1.73</td>
<td>&lt;.001</td>
<td>986</td>
</tr>
<tr>
<td>Middle School ELA</td>
<td>.3701</td>
<td>704391.42</td>
<td>260718.3</td>
<td>443673.2</td>
<td>1.63</td>
<td>&lt;.001</td>
<td>912</td>
</tr>
</tbody>
</table>

The results in table 12 are consistent with the prior two tables. All of the F statistics are statistically significant. In one case – middle school mathematics – there was more variation in value added ranking changes between schools than within them (SSB>SSW). The larger point of the three preceding tables is that there is significant variation in the changes in teachers’ value-added rankings between the schools in which they work. This suggests that there may be sorting of teachers into schools, not that schools themselves have an impact on how teachers are rated.\(^{34}\) Alternatively, it implies that schools (as the larger unit within which classrooms are contained) independently affect student performance, and therefore VA scores. We now move to see whether or not there are school characteristics which significantly impact teacher rankings above and beyond their own “abilities.”

Table 14 below presents the results of OLS regression models predicting teachers’ 2007 school year value-added rankings at different grade levels and in different subjects. Only two covariates are included in the models. The first is the teacher’s value added ranking from the 2005 school

\(^{34}\) Indeed Hanushek, Rivkin and Kain (2005) note that non-random assignment of teachers to schools is one of the key problems with estimating value-added.
year. The second is a “sheaf coefficient”\(^{35}\) composed of a group of school characteristics – the percentages of Black and Hispanic students in the school, the percentage of students eligible for Title I funding, the percentage of students classified as Special Education students, and a composite school environment score.

Table 14: Predicting 2007 teacher VA rank using 2005 teacher VA rank and a school-level
variables Sheaf coefficient, OLS Regression

<table>
<thead>
<tr>
<th></th>
<th>Elementary Math</th>
<th>Middle School Math</th>
<th>Elementary ELA</th>
<th>Middle School ELA</th>
</tr>
</thead>
<tbody>
<tr>
<td>VA_2005</td>
<td>.362 (.37)</td>
<td>.477 (.39)</td>
<td>.240 (.22)</td>
<td>.151 (.14)</td>
</tr>
<tr>
<td>School Sheaf</td>
<td>.999 (.09)</td>
<td>1.00 (.12)</td>
<td>1.00 (.14)</td>
<td>.999 (.18)</td>
</tr>
<tr>
<td>Model R(^2)</td>
<td>.148</td>
<td>.172</td>
<td>.070</td>
<td>.054</td>
</tr>
<tr>
<td>N</td>
<td>1,988</td>
<td>965</td>
<td>1,872</td>
<td>898</td>
</tr>
</tbody>
</table>

Note: Beta weights in parentheses. All coefficients are statistically significant at p<.05 or higher.

The results presented in the table are consistent across grade levels and subjects. A teacher’s ranking in 2005 is a statistically significant predictor of her ranking in 2007, which would be expected. But the sheaf coefficient of school characteristics is also significant in all four regression models; this suggests that these school characteristics do have significant impacts on teachers’ rankings. This is an important observation, especially given that, at least in New York City, the VA model is supposed to remove school level influences, including all those listed above (New York City Department of Education, 2011). Further, if we examine the beta weights connected to each coefficient we see that the relative importance of the sheaf coefficient is not inconsiderable in any case, and in one case – middle school English Language Arts – the school characteristics have more relative weight than does the teacher’s prior ranking.

---

\(^{35}\) Heise’s sheaf coefficient is a standardized multiple partial regression coefficient which treats two or more variables as if they were one, allowing researchers to summarize the effects of multiple indicators (Whitt, 1986).
This chapter attempted to demonstrate some of the empirical limitations of value-added models in identifying effective and ineffective teachers. We find that there is a significant instability in VA rankings using actual data from the NYC public school systems pilot of VA assessment. Specifically, VA rankings tend to “bounce” over time, suggesting that these measures are impacted both by variation in student and teacher performance from year to year. Given that standardized tests are imprecise measures of student learning, this should be expected.\(^{36}\) We further find that even within a single year, VA rankings can vary for a single teacher who teaches in two classrooms. This horizontal swaying suggests that VA is not capturing a stable trait of teaching, but rather heavily influenced by the mix of students in the classroom. Finally, we find that teacher VA rankings are persistently affected by school characteristics like race and socioeconomic composition and indicators of school environmental conditions. This finding calls into question the idea that VA holds teachers accountable only for factors they can control, a fundamental tenet of VA proponents.

Taken together with the previous chapter, which analyzed the theoretical assumptions of value-added alongside well-known practical realities of schooling, it is perhaps surprising to see that value-added and related types of assessment are being pushed as a cornerstone of education reform. This enthusiasm pervades through federal, state and local governments as well as private grant-giving foundations and many education advocacy groups. But logical consideration of evidence is not the only, or even the primary, engine of change. So we now turn to the history of education policy and assessment technology to understand the politics of reform over more than

\(^{36}\) Papay’s (2011) study of different VA using different test instruments, and the Gates’ Foundations Measures of Effective Teaching Project (Kane & Cantrell, 2010) both show that even in a single year, different assessments can generate different results.
a century of public schooling. We then proceed to trace the development of value-added assessment from its genesis in the 1970s through its ascendance in policy circles by the beginning of the 21st century.
Chapter 4: Value Added in Context: A Technological History of Educational Politics, or a Political History of Educational Technology

It’s perhaps unsurprising to conceive of the history of educational evaluation as bound up with “moral panics” over the state of schooling. From its outset, mass schooling in the United States has been described by some as grounded in the anxieties of the upper classes over the integration of working class and immigrant populations (Schultz, 1974). In the beginning of the 19th century in Massachusetts, high schools were proposed among a menu of options for the containment of restive youth which also included day-prisons (Katz M. B., 2001). These concerns were underlain by the prerogative of many industrialists that the working population of the future needed to be adequately disciplined for the new factory system of production (Nasaw, 1979).

But from the start, panics over schooling seem to be complemented by contemporary technologies of assessment and their inherent logics. The sources of anxiety about schooling are generally constant – problems of equity and quality are the regular themes. But in another way, the cart drags the horse as techniques create the content anew in successive eras. What the precise object of an educational moral panic will be depends on what and who new techniques are used to assess.

I propose here that educational reform movements, beginning with attempts at test-based accountability in the mid-19th century, through the “cult of efficiency” in the early 20th century and other movements of so-called administrative progressives, draw on technological

---

37 Stanley Cohen (2002, p. 1) defines a moral panic as when “A condition, episode, person, or group of persons comes to be defined as a threat to societal values or interests…presented in a stylized or stereotypical fashion by the mass media…”
innovations of the day for their legitimacy and capitalize on panics over perceived national decline. Furthermore, with the development of assessment techniques, the object of focus for reform efforts shifts. What Stanley Cohen (2003) calls the ‘folk devil’ – is influenced both by ideological narratives of the time and by the available scientific/technological state of the art.

1830-1850: Horace Mann, Samuel Gridley Howe, and the ‘Testing Wars’ of the mid-19th century

Perhaps the first instance of a technologically-informed panic over schooling involves the first comprehensive public school system in the United States. In 1845, the Boston Schools Committee, led by Samuel G. Howe and influenced heavily by Massachusetts school superintendent and fellow Whig reformer Horace Mann, administered the first system-wide assessment tests to students of the city’s public schools. These exams, and the report generated from them, had a profound effect on both the structure of schooling and the trajectory of testing in American education (Reese, 2013).

Before the creation of the public primary school system in 1789\(^{38}\), and for much of that system’s early history, Boston’s schools were locally funded and evaluated, usually through a system of ‘exhibitions,’ in which selected students would perform recitations of school material in public. A sort of talent show, students who gave the best performances won awards and honors. The purpose of these exhibitions was to garner financial support for the schools, particularly those

\(^{38}\) The city of Boston had five “free schools,” one for each of the city’s five main sections, since the early 18\(^{th}\) century. Those schools were regarded as among the best in the country, and their consistent funding by the communities reflected a widely shared belief in the need for schooling among the population. The Revolutionary war interrupted the continuity of the Boston free school system, as it did schools across the American territories. Massachusetts was the first state to pass a comprehensive schooling bill in 1789, and the city of Boston followed suit in the same year (Schultz, 1974, pp. 11-15).
which served poorer students and therefore relied on charity\textsuperscript{39}. Critics of the exhibition system suggested that it encouraged rote memorization and “spouting” instead of substantive learning, and that its competitive nature was morally suspect (Reese, 2013, pp. 21-23).

More importantly, some critics wondered whether the system promoted a false impression of school quality; after the formation of the public system, most Boston residents felt their schools were superior to those of other states and that their schools were generally improving each year. Boston’s grammar schools – at the time led by educator-administrators known as schoolmasters – operated autonomously under the traditional one-room schoolhouse model. Both Howe and Mann, borrowing from the school reform trends they saw first-hand in Europe, which Mann published to great fanfare and heavy criticism in his \textit{Seventh Report}, sought substantive reforms to the traditional operation of schools. Among these were the curtailment of corporal punishment; the introduction of an age-graded system with a standardized curriculum that did not favor any religious denomination or political party. Most importantly, reformers called for the centralization and professionalization of school leadership and evaluation. Each of these was at odds with the interests of the schoolmasters, who until then had relatively unchallenged reign over their institutions. But the masters had managed to avoid any major changes to the schooling system despite annual evaluations by state-appointed visiting committees (Schultz, 1974).

But the push for reform was abetted by two emerging trends in Europe espoused by both Horace Mann and Samuel Howe: phrenology and the nascent science of statistics. Phrenology, now discredited as a pseudoscience with racist overtones, was at the peak of its popularity in the mid-

\textsuperscript{39} Nasaw (1979) points out that the public school system in Boston, unlike those in Philadelphia and New York, served children from relatively diverse social backgrounds. Public schools required most families to pay tuition or ‘rates’ as they were then called, but poor families could apply for rate exemptions.
19th century and was seen by some as having socially progressive potential, specifically for education and corrections. Perhaps the most appealing feature of the Phrenology movement was its emphasis on empirical investigation and objective verification of mental traits, which until then had been left to the relatively subjective techniques of philosophy (Parssinen, 1974). This served as an important mark of rigor for the scientific community, but took on a quasi-religious fervor when Phrenology was presented to the larger society (pp. 8-9).

Howe’s Reading Committee tests and the resulting report were more directly influenced by early statistical analysis, which was also imported from Europe. The Manchester and London Statistical Societies were founded in 1833 and 1834 respectively, and some have directly suggested that they were founded to distance the collection of statistics from politics and establish it as a disinterested science (Cullen, 1975). Both Howe and Horace Mann were founding members of the Boston Statistical Society, the first in the United States, in 1839. Their fascination with statistics and belief that they provided objective facts about not only school performance, but social and political life more generally, served as the catalyst for using paper-

---

40 By 1823, England had twenty-four phrenological societies (Parssinen, 1974). By the time the chief spokesperson of Phrenology’s social applications, George Combe, came to lecture in the US in 1838, there were phrenological societies throughout the United States (Schwartz, 1952). It is also of considerable importance that Howe was one of Combe’s closest acquaintances and his greatest promoter in the US.

41 Popular education and expanded penal institutions were intertwined questions in 19th century Massachusetts. Both were seen as potential solutions to the problem of a rapidly growing working class population. Before becoming fully immersed in school reform, Horace Mann devoted a great deal of time to the cause of prison reform (Nasaw, 1979). Michael B. Katz (2001) points out that “Educational reform was one aspect of a widespread effort of government, both local and state, and private philanthropy to create a network of institutions capable of restraining the effects of the onset of large-scale manufacturing and increasing urbanism (p. 164).” Among these institutions were reform schools which more closely resembled prisons or forced labor camps, including the Boston Asylum for Indigent Boys which was modeled after the Mettray Penal colony in France. In *Discipline and Punish* (1977), Michel Foucault places the birth of the modern prison system with the opening of Mettray.
based examinations which would be assigned numerical grades which could be aggregated, averaged and illustrated in charts and tables (Reese, 2013, pp. 56-57).

The 1845 Boston Schools Committee, which first convened in January after a bitterly contested election, was divided between committee members aligned with the schoolmasters and those aligned with Mann, Howe and other early administrative progressives. When the committee first met, the members were mainly divided over the overuse of corporal punishment and the quality of schools serving African American children – the latter issue was important to many progressives, who were also abolitionists. Of course, these particular concerns came in the context larger discord over the competence of the school masters and the efficiency of the system in general, especially given the acrimony surrounding Mann’s *Seventh Report* (Schultz, 1974).

Testing and school comparison statistics were not on the agenda for the first half of the year, but would become a central point of contention thereafter.

The exams were administered in the last week of June, 1845. They were designed in private, away from the publicized debates over corporal punishment and school segregation; as such, the exam format and questions took the school masters by surprise. An indication of the examiners

\[\text{42 I borrow this term from David Labaree (2010) and David Tyack (1974), who use it to refer to one of two strands of the educational reform movement in the early 20th century. Educational progressivism, part of the larger progressive era, had two competing factions which espoused different solutions to the problem of inadequate public schools. One was the child-centered progressives, followers of John Dewey’s education philosophy who generally believed that education needed to be improved at the level of pedagogy by shifting away from rote instruction and toward what is now referred to as a constructivist approach. Administrative progressives, best represented by Edward L. Thorndike, believed the solution was to standardize the educational system and to differentiate instruction according to students’ proficiency. The history of schooling in the US suggests that the latter strand of progressivism was more successful and continues to carry the day. I use the term deliberately in reference to Mann and his adherents because I believe their influence created the strategic mold which Thorndike and others successfully deployed nearly a century later.}\]
derogatory purposes, vocabulary questions asked students to define obscure words such as ‘thanatopsis’; others asked trivial and ambiguous questions on geography. The exam’s primary authors were Howe on the Reading committee, and William Brigham, a fellow member of the Statistical Society, on the writing committee. Both men were committed to using the tests as evidence to push forward their favored organizational reforms in the schools. Since the tests were administered to all students in the first class of each school, and given the new format – written essays – and unannounced content, performance was predictably poor at many schools. But the effect of the examinations was amplified when the committee published comparative test-score data in tabular form in the committee report. Behind the scenes, Mann also “primed the public” for an attack on the school masters by having articles published and letters written in the newspapers which began to make a case for the necessity of favored school reforms (Reese, 2013, pp. 69-96).

The report of the reading committee was completed in early August of 1845, primarily by Howe, but with assistance from Mann, who was more proficient with mathematics and statistics. It painted a picture of schools which were not nearly as good as residents believed and of significant variation among the schools. Students across schools were classified as lacking knowledge in most subjects, a fact which the report linked both to prevailing instructional practices and to individual school masters. Inevitably, since the schools were ranked based on student grades, some schools were found to be at the bottom of the pack. The reading

43 In comparison to Howe’s 50-page report, Brigham wrote a much shorter 17-page report for the writing committee.
44 The only comparison with another town successfully completed by the Committee was an examination of an all-girls school in Roxbury, MA. The school scored higher on average than any of the Boston schools, but also had the second highest average age, second to another all-girls school within Boston (Reese, 2013, p. 138). Against the objections of some committee members, this comparison was heavily leveraged in the final report.
committee’s findings, including numerous statistical tables, were first presented as a working draft to the full committee on August 7th; the draft, which contained many mathematical errors, was only distributed to the entire committee three weeks later. Beyond the novelty of these data, however, the reports criticisms and recommendations were congruent with the basic demands of the reformers. Some of the results immediately reached the press along with the proceedings of the meetings, but the schoolmasters’ faction on the committee staunchly opposed printing and distribution of the full report beyond internal copies for committee members. The faction aligned with Howe recommended wide dissemination of the report, enough copies for every Boston family.

On September 11th, the committee voted to print and distribute 10,000 copies of the report, the first publication of its kind on issues of education. Because of its wide circulation, and because statistics and quantitative reasoning were a hot topic in the mid-19th century, the report received a great deal of attention in the press (Reese, 2013; Schultz, 1974). Based on the general presumption that the test score data provided factual evidence on the state of the city’s schools, reformers used the opening created to attack all aspects of the schools, not least of all the schoolmasters. Their opponents criticized both the tests themselves and the conditions of their administration. But the reformers achieved substantial success in their crusade, in large part due to the implications of the 1845 summer exams; by 1851 Boston’s schools were headed by a superintendent, the dual-headed system was ended, and written exams became the standard of evaluation displacing the exhibition system. Given the city’s pioneering nature and Mann’s high regard nationally, many of the reforms pioneered in Boston were adapted across the country. Dubious or not, this early use of an ascendant technology affected substantial change in early public school governance and practice.
1900-1920: Taylorism, Classroom Management, and the ‘Cult of Efficiency’

Decades later at the turn of the 20th century, evaluation technology would again play a major role in a frenzied period of education reform. The reformers would again come from the ranks of the social elite and from outside of educational quarters, but both the nature of the school system and thus the favored reforms changed substantially. In the era of the Boston controversy, migration and urban expansion were only beginning to create the need for mass schooling. But between 1870 and 1900, owing both to further population growth and to changes in laws regarding school attendance, the number of pupils in public schools more than doubled from fewer than 7 million to about 15.5 million; school expenditures had likewise increased from $63 to $215 million dollars. Particular progress had been made in terms of the enrollment of black children in schools: whereas less than 10 percent were enrolled in 1870, over 40 percent were in 1900; in the same period, the percentage of white children enrolled fluctuated between 50 and 60 percent (US Bureau of the Census, 1975).

This growth in both school attendance and expenses corresponded with a period of suspicion about large institutions which accompanied rapid industrialization, urbanization, and consolidation of wealth. Thus in the period beginning in 1900, American society began to react

---

45 Even in this period, population growth was driven by immigration from Europe, with an increasing proportion from poorer Eastern and Southern regions (US Bureau of the Census, 1975). Most of these migrants remained in East Coast US cities, assuring that the problems of perceived school underperformance would be concentrated in large urban systems.

46 By 1870, only three US territories – Massachusetts, Vermont and the District of Columbia – had enacted compulsory schooling legislation. But many such laws – with widely varying specifications – were passed in the next 30 years. The limitation of the legislation in many states was that they often lacked both the means to enforce attendance and sufficient school accommodations for all children (Katz M. S., 1976, pp. 17-20). Indeed, one of the concerns of reformers in this era was the need to improve compulsory attendance laws to prevent dropout and bolster retention, particularly in secondary school.
against various perceived excesses and abuses in industry and municipal government, including schools. Suspicion, and in some cases full-scale panic, was instigated by a number of journalists and publications collectively known today as muckrakers. But while the best exemplars of this tradition, including Ida Tarbell and Upton Sinclair, saw their painstaking journalistic efforts lead to positive changes in labor and housing regulation, environmental conservation, and anti-trust legislation; other figures of the muckraking tradition were neither so diligent in their investigative practices nor so scrupulous in their choices of targets (Callahan, 1962). Jal Mehta (2013) notes that perhaps the first critique of public schooling in the muckraking tradition came in 1892 when physician Joseph Mayer Rice lambasted rote pedagogical methods in a series of nine articles in the magazine *The Forum.*

Ironically, while much muckraking journalism and progressive reform efforts were being directed at the exploitation of labor and the environment by large industry, business values such as organization and efficiency were in a period of ascendancy among the American public. Corruption and vice among businessmen were viewed as episodic rather than systemic. The influence of business values on schooling was buttressed by powerful industrialists themselves. Andrew Carnegie frequently railed in speeches and in print against the impractical education – a reference to the traditional liberal arts curriculum – received by American schoolchildren.

---

47 The most sophisticated of the muckraking publications was *McClure’s* magazine. But in terms of popular influence, the most widely circulated magazines which contained critical stories about municipal school systems were *Ladies Home Journal* and *The Saturday Evening Post,* which each circulated more than 1 million copies per year at their peak (Callahan, 1962, p. 3).

48 Rice collected data from around 30,000 students, and found no relationship between time spent on spelling instruction and students’ later performance on spelling tests. His findings were not well received by school leaders, but such painstaking empirical data collection would become the trend in subsequent decades (US Congress Office of Technology Assessment, 1992). Rice would later go on to publish a collection of his essays in a volume titled *Scientific Management in Education* and, suggesting a pecuniary motivation apart from his muckraking work on the subject, a spelling textbook for use in elementary schools titled *The Rational Spelling Book.*
(Callahan, 1962, pp. 8-9). Henry Ford was no kinder to academic study, famously quipping in 1916 “I wouldn’t give a nickel for all the history in the world. It means nothing to me. History is more or less bunk.” John D. Rockefeller founded the General Education Board in 1901, whose first project was to create a program of ‘practical education’ for African Americans in the South – a program which was questioned by prominent Black intellectuals like W.E.B. DuBois and Carter G. Woodson (Marshall, 2011). More directly, many businesspeople would end up influencing municipal education policy by joining local school boards.

School administration was the primary target of education reform efforts. Both the criticisms of the liberal arts curriculum49 and the problem of increased costs were laid at the feet of professional schoolmen. Much of the discussion centered on the high schools, which were expanding at the most rapid pace, and which would be of most acute interest to business leaders and their government counterparts. Indeed, the key document which lays out the “social efficiency” vision of administrative reformers comes from the NEA Commission on the Reform of Secondary Education (Labaree, Someone Has To Fail: The Zero-Sum Game of Public Schooling, 2010, pp. 25-26).

The atmosphere of distrust and suspicion created in the early muckraking period reached a fever pitch regarding schools with the aid of proclaimed scientific research which deemed schools

49 The high school liberal arts curriculum was articulated by the preceding generation of common school reformers. Its last expression was to be found in the 1893 Report of the NEA Committee of Ten on Secondary School Studies – which consisted of six university presidents (including Harvard’s Charles Eliot), three high school principals, and the US Commissioner of Education. The report was very detailed in its recommendations for what should constitute a secondary school curriculum, addressing all major academic subjects (Dexter, 1906). Labaree (2010) notes that the report is very bureaucratic in its tone, arguing for a common program of study for all secondary students regardless of social class of college aspiration, but lacking the passionate appeal to values (e.g., democracy, citizenship) found in either Horace Mann’s common school reports or the Cardinal Principles report of 1918.
inefficient and with the prominence of a technique seen as a potential remedy. The research came in the form of a 1909 report by Leonard Ayres entitled *Laggards in Our Schools*\(^{50}\). Drawing on municipal school records, the report addressed two issues, students attending school below age-grade level (‘retardation’) and school dropout before eighth grade.\(^{51}\) Overall, the report deems one-third of all public school children retarded and cites an average dropout rate of fifty percent by the eighth grade (Ayres, 1909). Each could be considered obvious byproducts of the era: non-English speaking immigrants – and emancipated slaves who were denied literacy for centuries – would enter school below grade level as a matter of course, but were categorized as laggards for the simple fact that they were not in the appropriate grade. Similarly, dropout was often a symptom of the economic circumstances of families; most migrants were poor when they arrived and remained so thereafter, making work a necessity for many teenagers. Likewise, many African American families in the south were toiling under the highly exploitative sharecropping system. But Ayres’ report placed blame for these problems at the feet of the school systems:

“There is no teacher but will recognize the picture of this boy, and indeed, with some modifications, it fits many girls just as well. These are the boys and girls with whom this book deals. They are not the mentally deficient, exceptionally dull children. They constitute a large part of all of the school children in most, but not all, of our school

---

\(^{50}\) Leonard Ayres was a major figure in the history of Educational statistics, and an unapologetic promoter of education as a scientific discipline. In 1920 he was founding co-editor of in the *Journal of Education Research*, along with Lewis Terman and George Strayer. The journal was at the forefront of the movement to use scientific methods in the promotion of efficiency in education. He published numerous articles in the journal which dealt with topics including the computation of correlation and regression coefficients. His editorial work in the journal generally derided educators who opposed the use of science and testing as ‘the professionally moribund’ (Ayres, 1920a) who, ‘one wonders…notwithstanding the scholarly attainments which they sometimes possess, have not after all a fundamental intellectual defect’ (Ayres, 1920b).

\(^{51}\) A testament to its wide impact, *Laggards* was reprinted three times after its initial publication, including twice in the year 1910 and again in 1913. Ayres was a firm believer in the power of educational statistics – a later chapter of the book bears the title “Reform in and Through School Records.” He was also a believer in generating public attention for educational statistics; when he directed a 1916 school survey in Cleveland Ohio, educational news displaced war coverage on the front page of many Cleveland newspapers (Lagemann, 2002).
systems. These are the children that too many of our schools are confirming with the habit of failure (Ayres, 1909, p. 220).”

Ayres deemed these systems inefficient through a formula – the index of efficiency – which directly likened schools to factories, in which students were considered the raw material (Callahan, 1962, pp. 15-18).

The accountability technique that revealed both the problem and the solution was Frederick W. Taylor’s system of scientific management. Though he had been developing his ideas since the 1880s, Taylor’s most well-known elaboration of the concept came in eponymously titled Principles of Scientific Management which was published in 191152. Scientific management emerged as a result of Taylor’s experience as a gang manager at Midvale Steel Works, where he discovered the problem of ‘ordinary management’ and from his later work researching and experimenting with his famous ‘time and motion’ studies (Braverman, Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century, 1998). The four basic principles of scientific management were: 1) that managers must gather, study, tabulate, and codify all the traditional knowledge of the particular labor process into a set of the most efficient practices; 2) that managers must then match workers to jobs based on capability and train them to work for maximum output; 3) that work performance should be assiduously monitored to ensure application of efficient practices; and 4) that management should handle all planning and conception of work tasks in advance.

It is important to note the amount of research, and the resultant practices, required to practice scientific management. Taylor’s studies aimed at finding not only the most efficient methods for

---

52 Taylor’s first book on the subject, Shop Management, was published in 1903, and could have had a great deal of influence in the first decade of the 20th century among those beginning to criticize schools on the grounds of efficiency.
every function of the worker, but also the best tools and materials for the job. Requiring this level of rigor, his work at Midvale Steel took over 25 years. It is thus no accident that such a system coincided with the development of large production industries, since smaller firms would have had no way of paying for such intense studies. Even in the public consciousness, the Taylor system was associated most closely with the stopwatch (Callahan, 1962, p. 28). It is usually said that the original system of scientific management failed because it was both hated by labor and superseded by other schools of managerial thought. That is false both because scientific management succeeded in changing the basic assumptions of productive occupations (cf. (Braverman, Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century, 1998; Howard, 1985; Littler, 1978), and because in addition to resistance from workers, firms were unwilling to make the necessary research investments or to increase compensation for increased productivity.

Taylorism generated great, even mythical interest in the first decades of the twentieth century. The system became the subject of national attention in 1910 owing to congressional hearing of the Interstate Commerce Commission (ICC). A group of manufacturers moved to sue railroad owners over what were felt to be unfair increases in shipping rates. The railroads contended that increased wages for their employees made the rate increases necessary; lawyers for the plaintiffs called expert witnesses who testified that scientific management would make it possible to both raise wages and lower consumer prices without decreasing profits. The testimony was described as follows:
“Theirs was the firm faith of the apostles: it was a philosophy which worked, and they had the figures to show it (Callahan, 1962, p. 20).”

In the end, the commission sided with the Merchants, but not due to the evidence presented regarding the Taylor system, which was regarded by the commission as still in an experimental phase; the ruling was based on other more verifiable claims. Nonetheless, interest in scientific management – a term originally coined by one of the lawyers from the hearing, Louis Brandeis – grew throughout the 1910s. Hundreds of papers were published on the subject, applying the concept to nearly every facet of life. Efficiency societies formed and proliferated among the public, as did the ranks less rigorous yet ‘certified’ efficiency experts. An observer at the time remarked that ‘scientific management shingles have gone up all over the country’ without ‘standards or requirements, private or public… by which…the goats can be separated from the sheep.’

Given the existing suspicion of municipal institutions and corporations that emerged from the Muckraking periodicals and the growing fervor for all things efficient, especially scientific management, the school systems were an inevitable target for reforms. In the years immediately following the ICC hearing that brought Scientific Management to the fore, the primary venues for criticism were popular magazines. Sensational and polemical articles would stir anxiety among readers – a middle class clientele already stricken with the reform mindset – who would then express their frustrations at school board meetings and elections. *The Saturday Evening Post* featured stories with titles such as “Our Medieval High Schools,” and “Medieval Methods for Modern Children;” these stories criticized schools both for antiquated teaching methods and for

---

53 The use of expert witnesses and their often sensational testimony is a recurring story when we arrive at value-added assessment in the early 21st century.
administrative inefficiencies.\textsuperscript{54} \textit{Ladies Home Journal} was more virulent in its criticism of the schools, and more focused on efficiency. The editors cited the four-hundred million dollars spent annually on public schooling nationwide alongside dropout figures and literacy rates without any contextualization of either figure. Further stories described the public school system as:

“not only ineffective in its results but also actually harmful in that it throws every year ninety-three out of every one hundred children into the world of action absolutely unfitted for even the simplest tasks in life.”

“…it has grown antiquated and will not meet our present needs; it has indeed become a positive detriment…all traceable to its one great and crying defect – inefficiency (Callahan, 1962, pp. 51-52).”

Educational entities at various levels began responding to these early calls for efficiency first by putting scientific management on the agenda for discussion. The 1912 meeting of the National Education Association (NEA) featured panels dedicated to the idea of efficiency in schooling (Callahan, 1962, pp. 55-61)\textsuperscript{55}; the 1913 Superintendence Department meeting of the NEA was entirely devoted to the topic. At the federal level, the US Office of Education’s Commission on the Reorganization of Secondary Education report entitled \textit{The Cardinal Principles of Secondary Education} hinted at the need for the reorganization of secondary education based on the presence of students of varying aptitudes (Horn & Wliburn, 2013, p. 18). But as with Boston in the mid-19\textsuperscript{th} century, it would be outsiders to educational practice who would make the moral crusade for reform armed with an existing technology. Among the key figures were Ellwood Cubberly,

\textsuperscript{54} The effects of John Dewey’s child-centered pedagogy were clearly also being felt in articles such as these. But as Labaree (2010) points out, Dewey’s attempt to transform pedagogy was not as successful as the efforts of administrative progressives. This may be attributed to the durability of what Tyack and Cuban (1995) refer to as “the grammar of schooling” – that is day-to-day classroom practice.

\textsuperscript{55} The speakers on school efficiency who appeared at the 1912 NEA meeting ran the gamut between the prescient and the absurd. Some attempted to define the problems of applying scientific management to the education of children, while others gave attempted to define the “Material, Civic and Social dividends” to the public and “Janitor Efficiency.”
George Strayer, and Franklin Bobbitt; all three were from prominent schools of education and all focused their attention on the school systems in the largest American cities (Mehta, 2013, p. 45).

Some prominent educators and those directly involved with the schools made their suspicions of efficiency-driven reforms known from the outset. Education professor and former elementary school teacher William Bagley remarked in 1912 that “nostrums, panaceas and universal cure-alls in education are snares and delusions...we must give up the notion of solving all of our problems in a day.” Edward Hartwell, a statistician for the Boston schools, made a much more direct criticism, noting that “we should discriminate in our investigation and consideration of educational matter between methods and criteria that are applicable to living mechanisms and their activities and those which pertain to the realm of the inventor, the engineer, and the manufacturer (Callahan, 1962, p. 124).”

The attempt to reform schools in the mode of the Taylor system was fraught from the outset. Firstly, while the desired outputs of jobs such pig iron hauling, bricklaying and metal cutting were easily ascertained and checked for quality, the same was not the case for schools. Lacking a clearly defined output, school administrators who attempted to apply scientific management to their ‘school plants’ instead picked easily measureable objects of suspect validity which were either tangential to schooling, over-determined by out-of-school forces, or both. Among these were enrollment rates, average number of days attended by each student, and the number of students “inspired” to pursue higher levels of schooling\(^{56}\). Secondly, lacking a single decisive

\(^{56}\) These sorts of ‘efficiency’ mirror the problems and remedies laid out by Leonard Ayres. He suggests that the primary causes of ‘retardation’ were late entry and irregular attendance. His primary remedies for these problems and for dropout (‘elimination’ in his terminology) were more rigorous implementation of compulsory attendance laws (Ayres, Laggards in our schools; a study of retardation and elimination in city school systems, 1909, pp. 5-7). The other primary cause of retardation cited by Ayres was ‘physical defects’ among students, but these defects – the
output, few if any school systems attempted anything approaching a Time-and-Motion study to determine best practices. Nonetheless, the ethos of measurement-based school management remained strong. As a result, many reforms clothed existing concerns over school finances through novel ‘data-driven’ methods. In one case, Newton Massachusetts superintendent Frank Spaulding claimed to have made his schools more efficient by calculating the cost per hour of various school subjects to determine which were the most efficient uses of resources (Mehta, pp. 42-43). Spaulding’s approach was tantamount to “clothing business philosophy with the mantle of science.” (Callahan, p. 73)

As the push for school efficiency matured through the period of World War I and into the 1920s, it had two distinct threads, which Callahan refers to as the school surveyors and the efficiency testers (p. 99). The surveyors followed the lead provided by Leonard Ayres and defined problems of efficiency in terms of student flow, building use, and financial accounting more generally. The approach of Frank Spaulding described above exemplifies this version of school efficiency. Efficiency Testers more directly addressed questions of the educational process – teaching and learning – by relying on the developing field of standardized testing and psychometrics. Franklin Bobbitt emphasized the need to create external standards to determine whether, in a given school, teachers were capable and students were actually learning (Mehta, most common of which were poor vision and hearing, skin diseases, and defects of the nose and throat – would be today understood as consequences of poverty and child labor (Ayres, Laggards in our schools; a study of retardation and elimination in city school systems, p. 118).

57 Much breath and ink were wasted on the possibility. Franklin Bobbitt was probably the most articulate spokesman for applying Taylorism directly to the work of schooling, but even his writings only suggest the most basic experimental approaches to discovering best practices (Callahan, 1962, pp. 86-92). The emphasis throughout this period was devoted to rating teacher efficiency _ex post facto_ rather than making ‘capital investments’ to discover the most efficient methods.

58 Spaulding would subsequently lead school systems in Minneapolis and Cleveland, where he would institute data-driven administration practices (Lagemann, 2002, pp. 86-87).
As for the source of these external standards, Bobbitt suggested that ‘the commercial world’ ought to determine what should be taught in schools, since students were destined to become workers. Bobbitt further recommended that school systems create or adapt measurement scales for every product of schooling, which one city superintendent suggested would require school systems to maintain record-keeping systems many times more elaborate than even the largest corporations at the time (Callahan, pp. 83-86).

The importing of standards from business and from science was criticized by important educational figures, most notably John Dewey. In Education as Engineering he argues that efforts to found a science of education are doomed to failure without first pioneering new educational methods in the mode of science. Using an extensive analogy to bridge building, Dewey’s point is that if education is to be reinvented in any significant way, the reinvention will have to take place in practice before a new theory or science can be articulated. By contrast, if engineering or managerial methods are superimposed upon the educational sphere:

“Let it not be supposed that there is really any advance in the science of education merely because there is a technical improvement in the tools of managing an educational scheme conspicuous for its formation prior to the rise of science. Such ‘science only rationalizes old, customary education while improving it in minor details.” (Dewey, 2009, pp. 3-4)

The pressure for scientific management-based reforms of both sorts described above became institutionalized in a number of ways. Some large urban school systems established ‘efficiency

59 Among the architects of these original standardized tests in Mathematics, reading and other subjects was Edward L. Thorndike of Teacher’s College. Thorndike first developed his theories of learning through experiments with animals (Burke, 2013). His human intelligence tests, Alpha and Beta, were adapted by the US Military to classify soldiers during World War I, and were further transformed into the original Scholastic Aptitude Test (SAT) by Carl Campbell Brigham (Lemann, 1999).

60 Elsewhere, Dewey bemoaned the fact that although production industries use this prototype model of innovation regularly, schools lack the financial resources to do anything similar (Callahan, 1962, p. 94).
bureaus,’ internal offices which designed (or adapted) tests and scales to rate their teachers, students, and principals. Other municipalities hired efficiency experts, temporary consultants to conduct surveys and/or construct and administer rating instruments. For the most part, these surveying efforts were done in reaction to external organizations claiming inefficiency. But given budgetary constraints in many school systems, and the widely varying abilities of the ‘experts’ they hired, these surveys focused primarily on financial accounting version of efficiency, and only the most superficial aspects of the classroom situation. This was a marked change in the nature of the school survey, first initiated in Boston over a half century earlier. School administrators in this era much more closely resembled accountants than educators.

It is important to note that accountability in this period did not hold teachers solely responsible for student outcomes. Leonard Ayres’ *Laggards in Our Schools* emphasized the presence of mentally defective children in classrooms as a decisive factor in producing school inefficiency. In the language of moral panics, the victims would be gifted children whose learning was being held back by a tide of laggard youth now flooding the school system. Some attempted to develop measures of teacher efficiency based on student test scores (Conor, 1920; Elliot, 1915; Sears, 1921; Buckingham, 1920), while others stressed their primary function as measures of student abilities which could inform teaching and placement (Monroe, 1920; Willing, 1920). But the overwhelming emphasis of this early push for efficiency was on the classification of children as gifted, normal, or subnormal, which comes out of the emphasis on intelligence testing. In the

61 For example, the 1915 Cleveland school survey was a reaction to a claim by the Cleveland Engineering Society that schools were operating below 50 percent efficiency. In California and New York, surveys were initiated due to pressure from commercial interests and taxpayer groups (Callahan, 1962, pp. 115-116). Chauncey and Dobbin (1963) note that: ‘Early in the twenties, more than one-hundred bureaus of educational research were established.’

62 An illustrative quote from a 1921 *Journal of Educational Research* Editorial: “Conservation of power and reduction of waste is an important social ideal. A lot of the brain power in our schools
years 1920 and 1921, for example, one journal – *The Journal of Educational Research* – featured no fewer than thirty articles concerning the use of intelligence tests in schools compared to between five and ten on the subject of teacher efficiency. Another, *The School Review*, which was decidedly less interested in issues of measurement, had at least ten articles concerning intelligence tests in comparison with only two connected with teacher evaluation. Intelligence testing will thus be the focus of the next section of the chapter.

By the late 1920s, public attention had shifted away from school efficiency in the mode of the Taylor system. Indeed, Taylorism as a managerial theory was eventually overshadowed by Elton Mayo’s Human Relations approach. But the impact of the scientific management era in schools would last for decades. Ellwood Cubberly’s *Public School Administration*, first published in 1916, remained a primary text in schools of education even after his death in 1941, ensuring that school efficiency would remain a highly influential paradigm. Indeed, the prominence of school surveys in later decades’ efforts at school reform is a direct result of Cubberly’s influence. Thus the salient legacy is not that schools were forever changed into factory plants – schools that attempted such models did not last long – but rather that schooling came under the influence of the logic of business and industry without careful consideration of its utility for educational purposes (Callahan, 1962). Thus in the early 20th century, both the panic over schooling and the is daily going to waste because it is idling far below its normal working capacity. Genius which spends itself on inferior work is wasted. Our educational practice needs a severe earthquake that will shake us loose to the recognition of this waste.” The ideal is a recapitulation of Taylorism, but here the student is considered the worker, rather than the teacher. The general editor of the journal makes similar comments about the selection function of schools in a later editorial (Buckingham, *The School as a Selective Agency*, 1921)

63 Cubberly was a pioneer not only in school administration *per se*. He was also one of the most successful school surveyors of the era, conducting investigations in Baltimore, New York City, Oakland, Portland, and Salt Lake City (Eagle Shutt, nd). His history of schooling, elaborated in *Public Education in the United States* (Cubberly, 1919) offered a teleological and functionalist reading of American schooling which also remained dominant for decades.
promised solution were bound up with the assessment technology of the day; despite strong rhetorical resistance on the part of educators, public and interest group pressure fundamentally changed school administration.

1920-1950: Educational Testing Service, the SAT, and the rise of Mass Educational Testing

It was mentioned above that part of the push for educational efficiency was connected to the early development of psychometrics. Indeed, the entire concept of external educational standards requires that someone outside of the school design and apply measures of school performance. At the beginning, much of the impetus for developing tests of mental aptitude was clothed in a discourse of institutional efficiency. The focus then shifted to intelligence itself - invoking panicked tones about the state of the nation which would culminate in the passage of restrictive immigration legislation in 1924. As the fervor for mental testing and the labeling that came with it dissipated, mental testing underwent a change and became scholastic aptitude. It is this sequence of events to which we now turn.

While Horace Mann and Samuel Howe were among the first to administer externally created exams, standardization of assessment took a giant leap at the turn of the 20th century with the work of Alfred Binet in France and a group of US researchers who adapted his work. Binet, along with his colleague Theodore Simon, developed a series of tests and a resulting scale for the French school system to measure the intellectual capacities of French schoolchildren so that they could be placed in appropriate grade levels. The Binet-Simon scale was first released in 1905,

64 Attempts at developing ways of differentiating mental ability began as early as the late 19th century with the work of Sir Francis Galton and J. McKeen Cattell. The latter is credited with coining the term mental test in 1890 (US Congress Office of Technology Assessment, 1992).

65 It is important to note that Alfred Binet did not share the beliefs of those who adapted his scale in the United States. He was a disciple of John Stuart Mill, perhaps the most ardent early critic of
and was revised again in 1908 and 1911 after further field testing by the authors. Though they were not employed by the French government until after his death, Binet’s tests were put to more immediate use by two Americans. Henry Goddard, who is regarded by some as the father of intelligence testing in the United States, translated the Binet-Simon scale into English in 1908 and worked diligently to distribute copies to public schools. Lewis M. Terman revised the scale for American use in 1916; his version became known as the Stanford-Binet intelligence test and was the first to use the now-famous IQ scale (Plucker & Esping, 2013b; Plucker & Esping, 2013a). From the outset, both Goddard and Terman viewed intellect as an innate human capacity, and linked the quantification of intellect with notions of efficiency. Terman remarked in his 1916 book, *The Uses of Intelligence Tests* that:

> It is safe to predict that in the near future intelligence tests will bring tens of thousands of these high-grade defectives under the surveillance and protection of society. This will ultimately result in curtailing the reproduction of feeble-mindedness and in the elimination of an enormous amount of crime, pauperism, and industrial inefficiency (Terman, 1916, pp. 6-7).  

Goddard expressed similar sentiments in more strident terms in his *Human Efficiency and Levels of Intelligence*:

---

native intelligence. As such, Binet held that intelligence was, within limits, conditioned by environmental factors (US Congress Office of Technology Assessment, 1992, p. 113). Given his environmentalist position, he envisioned his tests to be used as full diagnostics which would not only classify children, but also be used to inform approaches for developing their capacities, thus increasing access to education rather than making it more exclusive. This is in contrast to American adaptors, especially H.H. Goddard and Lewis Terman, who used the Binet scale to advance eugenicist arguments (Minton, 1998). Lemann (1999, pp. 24-25) suggests that E.F. Lindquist’s Iowa Every-Pupil Testing Program was much closer to Binet’s intentions, as they were grounded in tests of achievement rather than aptitude.  

Illustrative of the connection between educational efficiency and intelligence testing, it is of note that Terman’s book was one in a series of education textbooks edited by Ellwood Cubberly. Terman was also a founding co-editor of *the Journal of Education Research* along with Leonard Ayres.
Testing of Intelligence is no longer an experiment or of doubted value. It is fast becoming an exact science. The facts revealed by the army tests cannot be ignored. Greater efficiency, we are always working for. Can these new facts be used to increase our efficiency? No question! We only await the Human Engineer who will undertake the work (Goddard, 1920, p. vii).

The ‘army tests’ referred to by Goddard were the ‘Alpha’ and ‘Beta’ tests used to screen draftees during World War I. Alpha was designed for recruits literate in English, while Beta was for illiterate or immigrant recruits. Goddard and Terman served on the design committee for the test, which was chaired by Harvard psychologist Major Robert Yerkes. The Alpha test, which was completed in the summer of 1917 and was largely based on the Stanford-Binet scale, was administered to over 1.5 million men by the end of the war. This constituted the largest instance of mass standardized testing at the time and provided invaluable data for intelligence test promoters to argue for their more general utility (McGuire, 1994; Armed Services Vocational Aptitude Battery, nd). It was also the first major application of the multiple-choice testing format, the development of which was credited to Arthur Otis, a student of Terman’s who created the first group intelligence tests (Chauncey & Dobbin, 1963, p. 6). The design committee emphasized the ease and efficiency (i.e., low labor costs) with which exams could be scored (US Congress Office of Technology Assessment, 1992, p. 124). Internally however, the value of the test was questionable: it correlated poorly with officers’ assessments of recruits’ capacities and was perceived by high-ranking military officials as making little contribution to the success of the war (Fancher, 1985, pp. 126-127).

---

67 Educational philosopher Jacques Barzun notes that “The vogue for this type of test began after the first world war” in his foreword to Banesh Hoffman’s The Tyranny of Testing (1963/2003, p. 7). E.L. Thorndike, a psychologist at Columbia’s Teachers College, was a primary constructor of early versions of the alpha, but his more complex view of intelligence conflicted with Robert Yerkes who believed intelligence was one quality that should be measured on one scale (Carson J. J., 2007).
The results of the alpha tests were kept secret until the end of the war. The National Academy of Arts and Sciences published the findings in 1921 under the title *Psychology Examining in the United States Army* (Army Alpha and Beta Tests, 2014). Like the school surveys conducted by Leonard Ayres, the test results provoked panic about the state of the nation’s intellect, though at first only indirectly about its schools. The official summary of findings stated that the average mental age of white draftees was approximately 13 years old and added that by this fact half of the white draftees would be classified as “morons” (Yerkes, 1921). Based on the results of the test, many authors bemoaned the menace of ‘racial impoverishment’ which they believed threatened to destroy American civilization (McDougall, 1921; Stoddard, 1922; Wiggam, 1922).

A review of Henry Goddard’s 1920 book stated that the results of the army tests placed the

\[\ldots\]

---

68 The significance of eugenics in the history of psychometrics is hard to overstate. Eugenics was introduced as a theory of social management by Sir Francis Galton, who also pioneered the concept of mental measurement in its service – though he failed to create an effective exam (Fancher, 1985). Most of the psychologists who helped to design the Army alpha exams – including Yerkes, Goddard, Termin and Bingham – subscribed to the idea native intelligence. Brigham’s *A Study of American Intelligence* loudly proclaimed that the nation was imperiled by a decline in intelligence owing to the presence of non-Nordic racial groups, a conclusion endorsed by Yerkes in his foreword (Brigham, 1922). Even though E.L. Thorndike differed with Yerkes on the nuances of mental measurement, he wrote unambiguously in support of native intelligence (Thorndike, 1913; 1940). The College Board’s decision to create a universal entry exam based on the alpha tests represented the interests of their elite clientele, which was concerned with increased applications from non-white – that is non-Nordic – students (Saretzky, 1982; Karabel, 2005). In 1948, reacting to an article which criticized the use of the SAT for college admissions, Educational Testing Service President Henry Chauncey reacted in a eugenicist tone: “If ability has any relation to success in life parents in upper socio-economic groups should have more ability than those in lower socio-economic groups. And if there is anything in heredity one would expect children of high socioeconomic group parents to have more ability than children of low socioeconomic group parents” (Lemann, 1999, p. 67). Even as late as 1963, Chauncey makes no distinction between intelligence testing and scholastic aptitude testing, though he distances himself from the native intelligence hypothesis, saying that although such a capacity is believed to exist, current instruments do not measure it (Chauncey & Dobbin, 1963, pp. 21-22).
science of mental testing beyond reproach before going on to advocate a national campaign of intelligence testing in the pursuit of greater national efficiency (Popenoe, 1921).

The Army alpha and the accompanying rising status of intelligence testing were not without strong criticism. Anthropologist Franz Boas challenged the racial impoverishment claim by stressing the importance of the social environment. He argued that in any supposed racial type are found a wide variety of individual traits conditioned by at least as much by social location as by anything inherited (Boas, 1922). As such, the findings of eugenics must be considered part of their political agenda rather than anything objective. Specifically addressing intelligence tests, he suggests they can provide no definitive answers in themselves:

…cultural environment is a most important factor in determining the results of the so-called intelligence tests. In fact, a careful examination shows of the tests shows clearly that in none of them has our cultural experience been eliminated…I suspect strongly that such influences can always be discovered and that it will be impossible to construct any test in which this element is so completely eliminated that we could consider the results as an expression of purely biologically determined factors. (Boas, 1931, pp. 12-13)

Walter Lippman’s series of essays in The New Republic not only criticized the fever-pitched interpretations of the alpha exam results, but also the arbitrary methods by which the tests were constructed (American Social History Productions, 2012). Lippman notes that Alfred Binet had shied away from attempting to define intelligence as such, opting to focus on the practical problem of identifying children who would need special attention in school, which he did by constructing arbitrary scales based on his own common sense attempts at defining aspects of mental ability (Lippman, Mental Age of Americans, 1922). By contrast, the Americans who adopted Binet’s approach – and worse still those who interpreted their research – treated exams

---

69 Lippman’s essays appeared between July and December of 1922 and were titled ‘Mental age of Americans,’ ‘Mystery of the A Men,’ ‘Reliability of Intelligence Tests,’ ‘Abuse of the tests,’ ‘Tests of hereditary intelligence,’ and ‘Future for the Tests.’
constructed by the same methods as if they measured an uncontested, heritable quality in humans (Lippman, The Abuse of the Tests, 1922). The numerical format in which the results appear – the quotient – further abstracts the results from their arbitrary basis and gives them an “illusion of constancy” which test promoters all-too-eagerly endorse. In the end, he attributes the eagerness of intelligence testers and their interpreters to overstate both the extent and the implications of the findings from the Army tests to a desire on their part for greater social power rather than any scientific commitment:

But of course no student of human motives will believe that this revival of predestination is due to a purely statistical illusion. He will say with Nietzsche that “every impulse is imperious, and, as such, attempts to philosophize.” And so behind the will to believe he will expect to find some manifestation of the will to power (Lippman, A Future for the Tests, 1922).

Despite the fact that later criticism showed the army tests to be roundly biased in favor of native-born whites – or perhaps because of it, given their eugenicist views – many psychologists who were involved with the testing program both touted and began adapting them for use in civilian life. Most notably, Robert Yerkes published the National Intelligence Test, which was widely administered. Walter Bingham, another design committee member, focused on using intelligence tests in industry and emphasized their role in increasing firms’ efficiency (Bingham, 1937). Lewis Terman, an educational psychologist with particular interest in gifted children who had

---

70 Even given his sharp criticisms, Lippman believed both: a) that the tests in use at the time would be superior to standard methods of classification used in schools (Lippman, The Reliability of Intelligence Tests, 1922) and b) that they could be usefully developed for a variety of applications, but for the pretensions that they measure intelligence (Lippman, A Future for the Tests, 1922).

71 In an article which defends the psychological research community of the 1920s against claims of racism, Snyderman and Herrnstein (1983) summarize criticisms of the Army tests from within and outside of academia. Specifically, they argue that the Army test results did not play a decisive role in the passage of the Immigration Act of 1924, as is argued by Stephen J. Gould (1981).
been working on the Stanford-Binet scale since 1916, used the results of the Army tests to promote widespread use of intelligence testing in schools."72

But in the long term, it was Princeton University professor Carl Campbell Brigham who had perhaps the greatest impact on standardized testing in education. Brigham was an assistant administering the Army Mental Tests during World War I. At first a strident eugenicist, Brigham contended that the tests proved the superiority of “Nordic” Europeans (Brigham, 1922). Upon returning to civilian life, he tried to make intelligence testing mandatory for admission to Princeton.73 By 1926, he had converted the alpha test into the first version of the Scholastic Aptitude Test (SAT). In the same year, after being appointed chairman of the testing committee at the College Board, Brigham administered his test to a group of college-bound high school graduates, just over 8,000 students. For the next decade, he was an influential member of the College Board and remained in charge of the exam. The early administrations of the SAT were not of consequence; the exam given to fewer than 20,000 students each year throughout the 1930s, and the College Board still made admissions decisions primarily based on its essay exams which had been in use since its founding 1901 (US Congress Office of Technology Assessment, 1992). But like earlier IQ tests, these initial pilots of the SAT provided later legitimation for the exam (Saretzky, 1982; Blackwell, 1999; Lemann, 1999).

72 Terman’s Stanford Achievement Test (which would now be considered a misnomer for what was essentially an intelligence test) popularized the use of the intelligence quotient (IQ) – defined then as mental age divided by chronological (Terman, 1916, p. 17). Terman borrowed the concept from German psychologist William Stern, who coined it in 1912 (Plucker & Esping, 2013).
73 At the same time, E.L. Thorndike was making similar attempts to introduce mental testing into the admissions criteria at Columbia and Cooper Union. Thorndike’s student, and Brigham’s fellow Alpha test assistant, Ben Wood would go on to develop the New York State Regents exams and the Graduate Record Examination (GRE).
The intervening years saw technological and political changes which would allow for the mass administration of college admissions testing. The technology was machine scoring. The SAT was a multiple choice test, but it still needed to be graded by hand. Ben Wood, another testing enthusiast and a student of E.L. Thorndike, had been working with the IBM Corporation to develop a machine to grade exams quickly and with fewer errors (Lemann, 1999, p. 35). They were unsuccessful, but did manage to purchase the rights to a scoring machine developed independently in 1933 by a Michigan high school teacher named Reynold Johnson. IBM released the first commercial scoring machine in 1937, and the SAT began using the machines in 1939 (Hadhazy, nd).

Concurrently in 1933, newly-appointed Harvard President James Conant, in collaboration with his new assistant Dean Henry Chauncey, was looking for a scholarship test to use to admit students from outside the university’s usual clientele of elite northeastern families. This goal led him to prefer the idea of an IQ or aptitude test over an achievement test, which he felt would only reinforce the existing prep school pipeline. Owing to Brigham’s claim that the tests predicted academic achievement and Chauncey’s promotion of the test, he chose the SAT and employed it to select ten scholarship students into the class of 1934 and again the following year. In 1937, with the cooperation of other testing advocates, Princeton and Yale each created similar programs which utilized Brigham’s SAT and a battery of multiple-choice subject tests designed by Ben Wood (Lemann, 1999, pp. 38-40).

From the outset, Conant’s aspirations were much larger than admitting a few Midwestern students to Harvard. On the heels of the 1937 administration of the scholarship exam, he proposed the idea of a national testing agency which would administer these exams as a national census of abilities for the purpose of discovering the country’s most talented minds. Taking his
inspiration from Frederick Jackson Turner, who had bemoaned the loss of America’s unique social levelling with the settlement of the western frontier; and from Thomas Jefferson’s notion of a natural aristocracy selected based on merit, Conant believed that a massive campaign of testing and sorting would turn back the tide of class stratification and national decline through the mechanisms of equal opportunity and enlightened leadership. Public education would provide the site for mass testing and consequently the place in which equal opportunity would be realized. And channeling those with the highest test scores into positions of social stewardship would ensure the best possible future for the nation.74

But Conant’s immediate plans for the SAT as a college entrance exam met opposition from what would seem an unlikely source. Carl Brigham had begun recanting his views on intelligence testing in the late 1920s during his tenure at the College Board. This was in no small part due to the zeal with which many intelligence testers had pursued their expansion, which Brigham felt came at the expense of rigorous evaluation of the tests themselves (Saretzky, 1982). Brigham’s criticisms of the tests were public and harsh, and his position at the College Board placed a major roadblock in front of Conant’s campaign.75 But Brigham’s failing health and early death in 1943

74 Conant’s ideas of discovering a Jeffersonian ‘natural aristocracy’ and fostering equal opportunity by means of the SAT was similar to remarks by Thorndike with reference to IQ tests: “It seems entirely safe to predict that the world will get better treatment by trusting its fortunes to its 95- or 99-percentile intelligences than it would get by itself. The argument for democracy is not that it gives power to all men without distinction, but that it gives greater freedom for ability and character to attain power” (Thorndike, 1920, p. 235).

75 Contrasting with his earlier book title “A Study of American Intelligence,” Brigham’s 1932 book on the SAT was titled “A Study of Error.” By 1934 he wrote that “The (intelligence) test scores very definitely are a composite including schooling, family background, familiarity with English, and everything else, relevant and irrelevant. The native intelligence hypothesis is dead.” (Saretzky, 1982, p. 16) His public comments on the potential effects of widespread intelligence testing were equally harsh, and suggested that such exams would have a deleterious effect on teaching and learning in schools (Lemann, 1999, pp. 40-41). Lee Cronbach differs about Brigham’s legacy, suggesting that his retractions were made on technical grounds rather than personal epiphany (Cronbach, 1975).
allowed it to move forward. The first step came with the onset of the Second World War; the College Board decided in 1941 that the essay exams would be too administratively complex to administer during war time, and opted to use the SAT instead – though this was still only of consequence at prestigious institutions. At the same time, Henry Chauncey was selected to adapt the SAT and administer it as a placement test for incoming recruits to the Army and Navy. The tests were given to over 300,000 high school students on April 2, 1943. This large scale, one-day administration of the Army Navy College Qualifying Tests proved that such tests could be efficiently and securely administered in a single day to vast numbers of students (Lemann, 1999; Horn & Wliburn, 2013). Over the course of the war, these exams were administered to roughly 10 million army recruits (US Congress Office of Technology Assessment, 1992). This brought the SAT to the larger population of students for the first time in its history.

The final decisive step in the ascendance of the SAT as a national exam came in 1947 when the College Board and the American Council on Education – another large national agency concerned with educational testing – were merged into one unit. The merger, opposed by both organizations for different reasons, was a direct result of Conant’s political power at the time. At the same time, the College Board handed over the task of administering its new primary exam, the SAT (the essay tests were never resumed after 1941) to a newly founded organization in Princeton, New Jersey – the Educational Testing Service (ETS). Henry Chauncey left Harvard to become the first ETS president in 1948.

---

76 The test was in fact a college admissions test; selected recruits were sent for advanced training at college campuses across the country. Similar tests were also given to enrolled college students for placement in the same programs.
James Conant saw college admissions testing as a means to a political end; Henry Chauncey saw testing as the end itself. Like the scientific managers before him, he was convinced that the rational information provided by test scores could replace the fraught decisions made based on tradition and sentiment. His immediate goal after the establishment of ETS was to administer a massive census of intellectual capacity to the entire country – an eerie similarity with avowed eugenicists twenty-five years earlier. Though this effort never materialized, ETS was provided a substantial government contract (nearly $1 million) to administer the Selective Service College Qualification Test (SSCQT) following the outbreak of the Korean War. Moreover, during his twenty-two-year tenure as its president, ETS created innumerable tests and expanded the administration of its best known device, the SAT, from around twenty thousand to over one million each year. In 1983, ETS was contracted to administer the National Assessment of Educational Progress (NAEP), a role which it continues to play today (National Center for Education Statistics, 2013).

More generally, the success of massive test administrations such as the Alpha and ANQT provided the model for states to begin large-scale testing of pupils as New York did with the Regents Exams. Perhaps the most notable statewide testing program was E.F. Lidquist’s Iowa Every Pupil Test of Basic Skills, which was first administered to elementary grade students in

77 The SSCQT exam – which like the ANCQT was based on the SAT - was one of two criteria which decided whether male college students would be allowed to continue attending college or be sent to the battlefield – the other was class rank in college. ETS administered the exam to over 500,000 students over the course of the war. Though about 60 percent of test-takers qualified for exemptions based on either or both criteria, the percentage scoring above the threshold score of 70 on the SSCQT varied greatly by region; reflecting the racial bias of the exams, those in the Northeast passed at a far higher rate (73%) than those in the Deep South (42%) (Freeman, 2009; Frusciano, 1983). Nonetheless, the debate over conscription at the time did not focus on the use intelligence tests in determining deferment, but on the idea of exempting college students in general.
From the outset, results of the Iowa tests were being used to evaluate the performance of students and schools. The exams became very commercially successful, and Lindquist eventually began marketing them to school systems outside of Iowa (US Congress Office of Technology Assessment, 1992, pp. 122-123). Another milestone achieved by Lindquist was a more advanced automated scoring machine, though this would not be realized until the 1950s. The use of mass-produced achievement tests like the ITBS in school systems paralleled and eventually overtook that of aptitude testing. In an interesting parallel with the application of scientific management techniques, test-makers pointed out that school systems and end-users (meaning principals and teachers) would ideally engage in intensive study of the tests they wished to use in their schools to make sure the tests align with the content being taught in their classrooms:

No one outside the school can do the job of stating what the school wants to measure…the job is wholly one for the teachers and administrators…Without it, achievement testing is scarcely more than random activity and contributes little to the work of the school…for the most part, the greatest benefit for the most students will result through the use of tests that have been chosen by committees of the students’ own teachers (Chauncey & Dobbin, 1963, pp. 57-64).

But the centralizing tendencies that came along with the introduction of these measurement tools made it unlikely that schools would get to choose the tests that best fit their instructional goals. Katz (1975) suggests that this process began in the late 19th century, and Wise (1979) argues that this centralization has a generally detrimental effect on educational quality. Further, as was the case with the proliferation of efficiency experts, many test publishers sprang up with the vogue

---

78 Because they did not come along with the sort of social labeling associated with aptitude tests, Lindquist’s achievement tests were not subject to the same level of criticism as were aptitude tests such as the SAT.

79 Between 1900 and 1932, there were more than three times as many published achievement tests available on the market as there were aptitude or mental capacity tests (US Congress Office of Technology Assessment, 1992).
for achievement testing and created what Chauncey calls “the quick-and-dirties that infest the darker part of the educational forest,” meaning hastily constructed exams which make sweeping and unsupported claims about their usefulness.

In the period encompassing the two World Wars, intelligence testing both unearthed the supposed problem of innate mental deficiency and became its apparent solution as a sorting mechanism at all levels of schooling. From the outset, mental testing as an educational policy tool had its critics. In addition to Walter Lippman’s criticisms noted above, William Bagley (1925) worried over the potential for mental testing to lead to a sort of social determinism – which presaged critiques of educational tracking decades later (Gamoran, 1992; Gamoran, 1987; Oakes, 1985). Banesh Hoffman made one of the most famous critiques of testing using the SAT as the representative in his 1962 book *The Tyranny of Testing*, arguing that professional testers relied on the perceived objectivity and numerical precision of their products to displace other forms of evaluation. Criticisms of the uses to which tests were put came from within the testing camp as well: ETS vice-president Henry Dyer said in 1971 that IQ tests and grade-equivalency scores used by schools often lead educators to “succumb to what Alfred North Whitehead called ‘the fallacy of misplaced concreteness,’” meaning that the numerical score is taken as an empirical fact rather than an imprecise indicator (Stevens, 1971).80

As with the Taylor-infused school efficiency movement, the focus of mass testing remained on groups of students and on large units such as school systems rather than individual schools or teachers. It would take further innovation to refine the targets of educational reformers. Furthermore, between their initial appearance and the end of World War II, testing proponents

---

80 Dyer went further and said that attempts to use such tests for ‘professional accountability’ were “absurd, wrong and misleading.”
shifted their emphasis from “classification for assignment” to “classification for allocating resources” in response to public criticism. Many – including Brigham, Terman, and Chauncey – also retreated (at least in public) from their explicitly eugenicist views. This is part of the reason why Lindquist’s achievement tests, and others like it, were able to ultimately gain more traction than many of the aptitude tests created in the same period. Though one cannot be sure, it is plausible that these were pragmatic shifts of position made to preserve the relevance of their tests and not genuine changes of heart or mind. Contemporary arguments for native intelligence, including those of Arthur Jensen and Herrnstein and Murray indicate the continuing appeal of the hypothesis.

But this is not the place to discuss the nature/nurture debate. What is salient is that large-scale standardized testing became a permanent feature of the American school system. The administration of the Army Alpha and later the SAT can be seen as blueprints for the program of national testing which begins in the 1970s. More directly, SAT scores themselves have been routinely editorialized as indicators of national educational progress from President Reagan’s wall charts through the present day. This sea change was in no small part due to the early and energetic efforts of test promoters who made outsized claims both about extant problems with schooling and the capacity of the tests to address them. These claims were buttressed by the scientific pedigrees of the testers, which allowed them to displace critics as luddites or traditionalists, and to gain influence among policymakers.


Around the same time as his critical essays on the Army Alpha exams were being published in *The New Republic*, Walter Lippman’s *Public Opinion* was released. Prophetic in many ways
throughout the book, he gives some attention to the role of social science in public policy. Aligning with the book’s general thesis that individuals only indirectly experience most the world (and therefore political life), he suggests that social science was poised to fulfill a great need “among the directing men of the Great Society.” He observes that at the time social science was still limited by the ad hoc ways in which their primary data were collected and by the inability to approximate experimental methods. But the increasing importance of a class of accountants, statisticians and other experts to people in power, he believed represented a fundamental shift analogous to that which happened to engineers and chemists – and therefore physical science generally – during industrialization. Increased contact between practical social scientists with decision-makers throughout society would result in more purposive data collection and steady advancement of theory (Lippman, 1997, pp. 235-237). This development would be further reinforced because:

…all large forms of association must, because of sheer practical difficulty, contain men who will come to see the need for an expert reporting of their environment…In the exchange of technic and result among expert staffs one can see, I think, the beginning of experimental method in social science. When each school district and budget, and health department, and factory, and tariff schedule, is the material of knowledge for every other, the number of comparable experiences begins to approach the dimensions of genuine experiment (Lippman, 1997, p. 237).

The American public school system expanded in both size and reach between 1950 and 1970. In 1950, there were 25 million public school students, with 83 percent of 14 to 17 year-olds attending (US Bureau of the Census, 1961). By 1970, those figures had climbed to 46.5 million and 94 percent respectively (US Bureau of the Census, 1970). A larger proportion of the population was completing primary and secondary education; the median number of years of school completed rose from 9.3 to 12 (US Department of Education, 1993). The growth in school attendance and completion was distributed across segments of the population, though the
gaps between races and regions of the country remained noticeable (Fischer & Hout, 2006, pp. 11-13). Mostly owing to the educational component of the GI Bill, postsecondary enrollment experienced perhaps the most remarkable growth, expanding more than three fold in the same period from 2.6 million to 8 million (US Congress Office of Technology Assessment, 1992). As in the time frames already discussed, immigration was the major force impacting K-12 school expansion during this period. But incorporation of previously excluded populations, specifically African Americans and rural populations, also contributed to this growth. As would be expected, the incorporation of new groups into the school system would again raise questions about school quality. Furthermore, given the achievement of near universal access to educational services, debates about social equality began in this period to shift from the issue of equity of access to the issue of differences in quality (Rivlin, 1971, p. 66).

Sociologist Arthur Wise also notes an important trend in the nature of educational policy in this period. Demands relating to education, both in terms of equality and overall quality, were directed at governments (state and federal) instead of at individual schools or school boards. Under increased pressure, legislative bodies attempted to address these grievances through legislation. Intentionally or not, these solutions changed both the operations and the public conception of the school system. Wise argues that while legislative solutions were the proper mechanism for addressing school equity issues, the same sorts of solutions are incapable of (and perhaps degrading to) addressing school effectiveness (Wise A., 1979).

There were two major catalysts for school-related panic between 1950 and 1970. The one that more immediately conjures up images of a panic was the Cold War – and more particularly the launch of the Sputnik satellite by the Soviet Union on October 4th, 1957. Though the United States successfully launched its own satellite less than four months later, the fact that its main
political and ideological competitor was the first to achieve such a milestone in space exploration created a sense of physical and ontological insecurity throughout the country. Critics – who were convinced of the failure of the schools long before *Sputnik* – capitalized on the opportunity it provided to advance their positions.

Historian and Constitutional scholar Arthur Bestor, who published a manifesto on the question of school quality in 1953 titled *Educational Wastelands*, was interviewed in January of 1958 by *US News and World Report* on the question “What went wrong with US schools?” His argument was in that progressive school reforms – implemented by a group of “professional educationists” – had failed to produce the needed mathematics and science talent for the United States to be competitive with the Soviet Union (Tyack & Cuban, 1995, pp. 52-53).81 *Life* magazine further stoked panic over the state of the nation’s schools with a five-part series titled “Crisis in Education.” Accompanying essays articulating national educational decline were photo spreads comparing two high school students – Alexei Kutzkov (Russian) and Stephan Lapekas (American) – in a way that depicted Soviet education as academically serious and American education as diluted by socialization and too concerned with trivial matters (Bracey, 2007).82 Another influential trumpeter of educational decline following the launch of *Sputnik* was Admiral Hyman Rickover, who is also credited with the development of the first nuclear submarine. The admiral published a series of books beginning with *Education and Freedom* in

---

81 Interestingly, University of Illinois Education professor Harold Hand calls Bestor’s take on American schooling a “Devil Theory” (Hand, 1958) in an argument which predates the theory of “folk devils and moral panics” outlined by sociologist Stanley Cohen in the 1970s (Cohen S., 2002). Hand notes that Bestor’s argument about the effects of progressive education on the American space program would have been minimal at best – since recent high school students would not have been working on the space program – and that his indictment of an ill-defined group of educational decision-makers did not withstand investigation.

82 The first installment of the ‘Crisis’ series appeared in the March 24th, 1958 issue, and featured an essay by novelist Sloan Wilson titled *It’s Time to Close Our Carnival.*
1959, and continued with two comparative studies of US schools with those in England and Switzerland (Ducros, nd). Each volume leveled similar criticisms about the content of American schooling, charging that it has been leveled toward the less capable students by emphasizing “life adjustment education” rather than fundamental content, particularly math and science.\footnote{Based on an interpretation of the work of John Dewey and other child-centered progressives, ‘life adjustment education’ aimed to increase schools’ emphasis on the practical applications of learning (McFarland, 1954). The move toward this practical-skills curriculum began to take hold in 1945 in response primarily to a high dropout rate and a generally low college-going rate among US high school students. By the time Sputnik was launched, critics of life adjustment had already gained traction in undermining it (Breault, 2009). Lawrence Cremin (1961) regards the Life Adjustment movement as a key misstep of educational progressives.}

The major response to Sputnik came in the form of the National Defense Education Act (NDEA), which was signed into law in September of 1958. The majority of the law’s impacts were at the postsecondary level, but it did turn a new page in the story of education reform at the K-12 level, as federal funding was used to leverage changes in state education policies. A multitude of new subject-area initiatives were started following the NDEA, resulting in new textbooks and curriculum protocols (Buxton, 2009; Rudolph, 2002). Notably, in 1959 the US Office of Education fully funded an already existing study by the University of Pittsburgh and the American Institutes of Research, known as Project TALENT, which was charged with assessing the quality of education in US high schools.

While the Cold War and the panic over Sputnik initiated a struggle over US high school curriculum, the nascent Civil Rights movement simultaneously drove another set of reforms in the school system aimed at desegregation and school equity. Mandatory desegregation of public schools began with the Brown v. Board of Education decision in 1954. As schools attempted to integrate in the years that followed, focus shifted to the tasks of educating diverse populations...
and the problems of school equity (Weiner, 1993; Mehta, 2013). On the one hand, some education scholars used terms like ‘culturally deprived’ and ‘disadvantaged’ to refer to students from working class and poor backgrounds, whose values and capacities did not accord with those purported in school (Reissman, 1962; Passow, Goldberg, & Tannenbaum, 1967). These terms and the approaches to schooling associated with them became controversial as a result of the “culture of poverty” thesis advanced by Daniel Patrick Moynihan (1965) in The Negro Family: A Case for National Action. Other writers focused on the drastic inequalities between the schools serving black and white students (Kozol, 1967; Rosenthal & Jacobson, 1968). Stanley Aronowitz (2000, pp. 114-115) suggests that increased investment in education coincided not only with fears of Soviet competition, but also with trepidation about rising rates of crime following the economic boom of the postwar years. Elsewhere, Moynihan (1970) suggests that concerns over juvenile delinquency – especially among urban Blacks and Puerto Ricans – were reigniting a fear of proletarian violence that was common in the 19th century.84 Linking the threads of the Cold War and fears related to minority communities, some have suggested that the War on Poverty programs of the 1960s were an attempt to buttress the poorest American communities so that they wouldn’t provide fuel for Soviet propaganda abroad (Balogh, 2004).

Bolstered by the broader civil rights movement and Cold War fears, and thus incorporated as part of Lyndon Johnson’s ‘War on Poverty’ agenda, the question of racial inequality in education became more pressing than that of educational standards. The introduction to the Elementary and Secondary Education Act of 1965 makes clear that the intention of the legislation is to “…

84 Moynihan notes that the rates of crime and poverty in New York in the late 1950s and early 1960s were lower than those in the late nineteenth and early twentieth centuries, but that expectations had changed such that relative to what people expected, there was a perception of a real and growing problem of juvenile delinquency (Moynihan, 1970, pp. 43-46).
contribute particularly to meeting the special educational needs of educationally deprived children.” The fact that Project TALENT had far less public impact than the government report on ‘Equality of Educational Opportunity’ (the Coleman Report) gives evidence to which issues dominated the period.

But as the story below suggests, Cold War rhetoric and the techniques developed to prepare for both armed and ideological warfare would find their way into the sphere of domestic issues, including education. Part of the strength of this school equity agenda in the period came from a technical innovation which had been developing in the military since World War II: Systems Analysis.

Systems Analysis (SA) encompasses a wide variety of mathematical and statistical techniques, including multivariate modeling and simulations, but is generally oriented toward generating ‘optimal’ solutions to administrative problems by defining and quantifying the inputs and outputs of a given situation.\(^{85}\) The story of Systems Analysis begins with Operations Research (OR). As a formal discipline, OR originated in Great Britain as the country prepared to deploy radar systems to combat aerial bombing campaigns by the German Luftwaffe. This original application concerned the best uses of machines, although it was quickly adapted to the analysis of human systems in the military context (INFORMS, nd).

\(^{85}\) A report from the RAND corporation describes Systems Analysis succinctly: “By a ‘system we mean a set of interrelated factors that are used together to produce an output...All of these systems produce some sort of output...[which] are typically difficult to quantify...[and] also have costs associated with them...[and] have various ways of combining the elements or inputs in order to produce outputs...In a successful systems analysis, the analyst can vary the inputs, note the effect on both cost and output, and then decide that one system is better than another...if you tell him what output you want, he can tell you how to get it at minimum cost (Kershaw & McKean, 1959).”
Branches of the US military, primarily the Navy and Air Force, began separately using OR techniques to inform bombing campaigns and anti-submarine warfare in 1942. Among those contracted to teach these techniques in the military was former accounting professor and future Defense Secretary Robert McNamara (Mehta, 2013, p. 69). In 1945, following the conclusion of the war, the US government established Project Research and Development (RAND) under a contract to the Douglas Aircraft company, with the intention of applying OR-style techniques to identify and solve long-term military problems (Gass & Assad, 2005). The project’s reputation and staff grew rapidly over the next three years. In 1948, Project RAND’s government contract was transferred from Douglas Aircraft to the newly established nonprofit RAND Corporation (RAND Corporation, nd). Among RAND’s first objectives as an independent entity was to refine OR to deal with military problems which involved more uncertainty than radar or bombing. The stated goal was to create a comprehensive ‘science of war’ which considered a variety of outcomes, and Systems Analysis was the methodological basis. But the techniques proved incapable of this optimization given the various contingencies of war. Though RAND continued to be influential in military planning during the 1950s and 1960s, Systems Analysis and its applications to social welfare research are considered its most far-reaching contribution to American policy-making (Jardini, 2000). But the key to understanding RAND’s transition to domestic and civil affairs lies in the discipline of economics.

Economics was a small branch at the RAND Corporation in its early years. But by the mid-1950s, it came to occupy a central place in the organization because economists’ notion of cost-

---

86 At the same time as OR methods came into use in the US military, sociological methods were also being utilized for the first time in the Army. Samuel Stouffer’s study of army life, which made use of developing survey research methods, was commissioned around the same time (Bulmer, 1996). One could characterize the 1960s as a period when sociology and economics competed for influence on policymakers, with economists emerging as the victors.
effectiveness became the outcome of choice for systems analysis techniques. Cost-effectiveness was not only an appealing goal for the military; it also suggested a way to apply SA other problems (Jardini, 1996, pp. 107-113). RAND Corporation’s first statement on the applicability of Systems Analysis to education came in a 1959 memorandum to the Ford Foundation – which provided them a small grant of $35,000 – outlining potential difficulties of studying schooling in terms of the relationship between inputs and outputs. The lack of an adequate definition of school output should, the validity and reliability of tests as measures, and the problem of finding school systems which are ‘similar in all respects except one’ are among the problems discussed by the authors (Kershaw & McKean, 1959). The paper presages much of what James S. Coleman found in his report – also discussed below – including the relatively small role played by schools in educational outcomes and the important role of “peer effects” (pp. 20-27). Ultimately, however, the authors conclude that even if the prospect of comparing school systems is a ‘long shot,’ it is the “only hope of providing a systematically better basis for intelligent choices in education.” As Jardini (1996, pp. 359-360) points out, this study was RAND’s first foray into the field of domestic policy research, though they would not put the full force of their organization into this line of inquiry until the late 1960s.87

Prior to Systems Analysis, government interest in using economists’ tools for policy planning had already made an impact. H. Thomas James (1968, pp. 4-6) notes in the *The New Cult of

87 RAND Corporation’s most high-profile attempt to move in this direction was an attempt to create a domestic policy think tank with the same relationship to the federal government’s department Housing and Urban Development that it had with the Air Force. This attempt failed, as the Ford Foundation backed other interests in the founding of the Urban Institute, but RAND was still able to gain footholds in social research through a program on Middle East studies (1967-1973) and by establishing an independent institute in New York City (1968-1975). Even during this early period, the use of independent consultants as evaluators of city programs led to significant criticism of the mayor’s office (Jardini, 1996).
Efficiency and Education that the work of the 1947 Hoover Commission exemplified a growing importance of economic analysis and a renewed interest in efficiency in the federal government. Following the commission’s 1949 Report, President Truman submitted over thirty reorganization plans for various executive offices. President Eisenhower also promoted efficiency-driven reforms, including the National Defense Reorganization Act of 1958. But this reorganization was only put into practice in 1960 under the new Secretary of Defense.

When Robert McNamara became Defense Secretary, the influence of Systems Analysis on US national policy increased dramatically. Between 1961 and 1963 – what was known as the “McNamara revolution” in Office of the Secretary of Defense – military concerns of all sorts were redefined as economic problems reducible to the efficient allocation of resources.

McNamara, who had been an accountant for the Ford Motor Company immediately after the war, sought to implement in the Department of Defense the top-down management techniques he had implemented at Ford, collectively known as “statistical control.” He found willing colleagues at the RAND corporation, who shared his mindset with regard to efficiency and management and who were growing frustrated in their partnership with the US Air Force (Jardini, 1996). McNamara brought many former RAND employees into his office; his hope was to change DoD operations without changing the organizational structure of the military.

---

88 Statistical (Quality) Control methods originated with the work of Walter Shewhart in the 1920s. They utilized sampling to ensure consistent quality of manufactured goods and to find variations in quality which could be remedied by improving aspects of the production process (Deming W. E., 1982). The techniques were adapted for use during World War II largely based the efforts of W. Edwards Deming, who proceeded to have a long career as a business consultant especially in post-war Japan. Perhaps ironically, Deming opposed setting numerical goals for employees.
The outcome of the reforms of the Department of Defense under Secretary McNamara was Planning, Programming and Budgeting, or PPBS (West, 2011). The principles were based largely on the work of two RAND economists, Charles Hitch and Roland McKean’s, in The Economics of Defense in the Nuclear Age. As with Taylorist efficiency experts in the early 20th century, Systems Analysts made the analogy between their intended object, the conduct of warfare, and the efficient manufacture of raw materials:

“…the problem of combining limited quantities of missiles, crews, bases, and maintenance facilities to ‘produce’ a strategic air force that will maximize deterrence of enemy attack is just as much a problem in economics (although in some respects a harder one) as the problem of combining limited quantities of coke, ore, scrap, blast furnaces, and mill facilities to produce steel in such a way as to maximize profits” (Hitch & McKean, 1960).

PPBS became the standard for military planning at the highest levels shortly after McNamara’s appointment as Secretary. Charles Hitch was among the lead architects of the transformation of the DoD to the system. Its implementation had disastrous consequences for the conduct of the Vietnam War, during which goals were defined in quantitative terms. Donald Campbell (1975) retells the notorious ‘body counts’ as an example of his famous law about the corrosive effects that follow from the ‘worship of a quantitative indicator.’ McNamara later expressed regret over the excessive faith he had placed in rational management on camera in The Fog of War (Mehta, 2013, pp. 70-71). Perhaps the film would have been more aptly titled “The Fog of Statistics.” But at the time, enthusiasm for PPBS and Systems Analysis generally was running high in Washington.

The migration of PPBS to domestic policy took place amid political controversy over the Great Society programs. When the initial attempt to implement the 1964 Economic Opportunity Act through direct funding to locally-administered Community Action Programs proved divisive, the
Johnson administration reorganized the program along the PPBS model (Jardini, 1996, pp. 326-340). In retrospect, the central problem with the Great Society programs was that, given a stretched federal budget due to the mounting costs of the war in Vietnam, the Office of Economic Opportunity was never able to implement its programs at an appropriate level to meet the problems of urban poverty (Walinsky, 1969). Furthermore, the collapse of these programs coincided with a backlash against such policies by working class whites who felt their benefits went overwhelmingly to African Americans, and who were being actively courted by Johnson’s challenger, George Wallace\(^89\) (Bush M., 2004; Kaufmann, 2004). Nevertheless, its failure was attributed to the community action framework (Moynihan, 1970), which was largely phased out when the law came up for renewal in 1967.

McNamara and systems analysis were influential among federal officials in charge of Great Society programs. The director of the Office of Economic Opportunity, Sargent Shriver, was in close communication with McNamara during the planning of the Economic Opportunity Act. Other McNamara supporters included economists in the Bureau of the Budget who had worked with Charles Hitch in the implementation of PPBS at the Department of Defense. When Shriver was relieved of his position at the OEO in 1968, he was replaced by a former Budget Bureau analyst (Moynihan, 1970, pp. 4-5). In August of 1965, President Johnson signed a directive ordering the implementation of the PPBS across all federal cabinet departments. Though outcomes in the domestic sphere were much more difficult to define, RAND employees and alumni were hired to work in various federal offices to manage the transition to the Systems

---

\(^89\) Additionally, many whites in both the north and the south began to respond negatively to a change in the character of the Civil Rights movement, which up until 1964 focused primarily on issues of integration, but shifted focus toward Black Power and Black Nationalism in the late 1960s (Bush R., 1999).
Analysis approach (Jardini, 1996, pp. 341-343). Moynihan later remarked that PPBS “colonized the federal establishment” (Mehta, 2013, p. 71). In the same year, the original Elementary and Secondary Education Act was passed, which allocated $3 billion for federal education funding.

The vast expansion of systems analysis techniques at the federal level filtered down to state governments as well. Alice Rivlin’s *Systematic Thinking for Social Action* crystallizes the PPBS approach outside of the military, and shows the degree to which the logic and practice of systems analysis were accepted in the halls of government. In her own words, “Missionaries no longer need to be sent to convert the heathen to the virtues of systems analysis” (Rivlin, 1971, p. 6).

Other authors spoke about the potential of Systems Analysis in tones reminiscent of efficiency experts and intelligence testers:

“(Systems Analysis) enables one to raise probing questions in a universal language. Systematic thinking is logical thinking…The range of potential applications…is nearly unlimited…Its major virtue is the enhancement of human judgment.” (Hartley, 1968, p. 28)

“The argument for systematic experimentation is straightforward: Information necessary to improve the effectiveness of social services is impossible to obtain in any other way.” (Rivlin, 1971, p. 108)

With the application of PPBS and cost-benefit evaluations across the domestic policy spectrum, many compensatory programs were found to be ineffective or too expensive (Rossi & Freeman, 1993). With regards to compensatory education programs, results were particularly dismal;

---

90 In 1968, Rivlin was appointed Assistant Secretary of the Department of Health, Education and Welfare in charge of planning and evaluation. The Secretary at the time was William Gorham, who began his career at the RAND Corporation, and would go on to found the Urban Institute based on a federal contract to evaluate President Johnson’s social welfare programs. In this way, the same procedure that gave rise to RAND in the military brought economics-driven policymaking to civilian life. *Systematic Thinking* was published based on a series of lectures given by Rivlin to the Brookings Institution in 1970. The two previous speakers in the same series were Charles Hitch and Charles Schultze; the latter was the director of the Bureau of the Budget in the Johnson Administration.
Levin (1972) notes that of 1,200 such programs evaluated by the Office of Education, only 10 were shown to be successful.

Systems Analysis in education saw its most notable exponent in 1966 with the publication of a report commissioned by the US Office of Education. The report – titled “Equality of Educational Opportunity” – was mandated by the Civil Rights Act of 1964, and quickly became associated with its primary author, sociologist James S. Coleman, earning the nickname ‘The Coleman Report.’

Based on a massive study involving nearly 650,000 students and teachers in over 3,000 schools, it examined the impact of school segregation in terms of “inputs” and “outputs.” The report found that most students attended racially segregated schools; white students attended the most segregated schools, while Black students were the most segregated among minority groups. Schools with mostly black students had fewer resources and were staffed by less credentialed and less educated teachers. Racial gaps in achievement – measured by ETS-created standardized tests of mathematics and verbal ability – were significant and increased with age.

But the most notable finding regarded the relationship of school characteristics (inputs) to

---

91 *The Coleman Report* was not the first large-scale attempt to measure educational inputs and outputs and relate the two. Researchers at the American Institutes for Research (AIR) and the University of Pittsburgh, led by John C. Flanagan, conducted a national survey of high school students called Project TALENT. But as Rivlin notes, the study focused only on high school students, did not collect race data, and did not lend itself to socioeconomic comparisons (1971, p. 13). Flanagan was associated with Air Force aptitude testing during World War II and founded AIR shortly after. The study was fully funded by the US Office of Education only after the Soviet launch of Sputnik.

92 An important link to the previous section is that although the principal investigators, Coleman and Ernest Campbell, were both university academics, the survey coding and much of the analysis were conducted at the offices of Educational Testing Service. This is consistent with the increasing role of private research firms in conducting early social policy research in the 1960s (Bulmer, 1996). Perhaps more relevant, the director of the National Center for Education Statistics, Alexander Mood, had been a researcher with RAND Corporation from 1950 to 1955 before founding his own OR firm. Mood supervised the Coleman Report and played an important role in computerizing the statistical arm of NCES (INFORMS, nd; Mosteller, 2009).
measured achievement (outputs): the report found that students’ family backgrounds and the characteristics of their peers in school played far greater roles in determining achievement than did any measure of school quality. The implication of this finding was that efforts to improve student achievement by increasing school resources – one of the intentions of the Johnson Administration – would do little in terms of educational output. For our purposes what is most important here is that the school was conceptualized in the language of the systems analyst.

The general finding in Coleman that students’ outcomes varied significantly by school, ignoring the report’s important caveat regarding the broader social context, raised the issues of school performance and a new specter: accountability. By 1971, states from New York to Kansas to California were calling for schools to be held responsible for student performance. Fred Hechinger called accountability the ‘great current controversy across the nation’ (Hechinger, 1971).

President Nixon’s Message on Education Reform included a section on “New Measurements of Achievement” which began with the question “What makes a good school?”;

---

93 Beyond the reports’ political impact, its repercussions for the community of researchers concerned with education were enormous. The shift in focus from educational inputs – facilities and resources material and human – to outputs like test scores marked a decisive change in the nature of discussions about what matters in schooling. In another chapter, I examine the evolution of Value-Added Assessment techniques, a story which begins with responses to and critiques of the Coleman Report.

94 There is a complexity which should be mentioned at this point. On the one hand, the concerns over school equity indeed had factual grounding; Coleman brought a large amount of data to bear on a problem which was likely obvious to casual observers and social scientists alike. The point about moral panics and moral crusades is not that they create problems out of thin air, but rather that they mark a focal point for public outcry. As Tyack and Cuban (1995) and Labaree (2010) point out, the educational system has frequently been visited with reformers using moral outrage as a tool. What I attempt to add to this discussion is the observation that the chosen target depends heavily on the state of social science and its techniques.

95 Ironically, in the same article, Hechinger quotes former RAND economist and Systems Analysis pioneer Charles Hitch – then serving as University of California President – in opposition to Governor Ronald Reagan’s call to impose accountability metrics on the UC system. Such measures, Hitch said, ‘would reduce our flexibility to make best use of our resources.’ It’s rough when the axe swings in the other direction.
therein, he advocated a shift in focus from educational ‘inputs’ to ‘outputs’, emphasized the need to develop ‘more sensitive measurements’ of school quality, and called for accountability, though in vague terms (Nixon, 1970).

As with Systems Analysis in general, its application to school accountability was trumpeted by supporters as the antidote for all the ills of public education. “Educational Engineering,” as Assistant Education Secretary Leon Lessinger called it, would facilitate educational accountability for results. Without specifics, he suggested that accountability in itself had the capacity to affect change in every aspect of education from curriculum to student boredom to school segregation (Lessinger, 1970; 1971).96

Some of the first experiments with school accountability went under the heading of performance contracting, in which outside firms were brought in to teach underperforming students and were paid on the basis of students’ growth on standardized achievement tests. Beginning with a program in Texarkana, Texas in 1969, by 1972 there were around 150 performance contracts in effect throughout the country. Teachers’ and Administrators’ unions reacted harshly to the introduction of performance contracts; “Hucksters in Our Schools!” was the title of a 1970 article in American Teacher addressing the issue; teachers in Gary, Indiana threatened to strike in the same year over a performance contract in that city. Some test-makers also reacted unfavorably toward this use of standardized tests in performance contracting, as it became clear that the tests were directing instruction (Mecklenberger, 1972; Ascher, 1996). The push for performance accountability

---

96 Lessinger was perhaps the most vocal spokesperson for educational accountability of the era. His 1970 book, titled *Every Kid a Winner: Accountability in Education* was described as the “bible of accountability,” and many authors followed suit with more specific accountability titles in the next five years.
contracting ultimately ended in controversies over contractors’ ‘teaching to the test’ and the lackluster results of program evaluations (Rosenthal J., 1972).97

The first major municipal effort at school accountability was initiated in New York City. Mayor John Lindsay was enamored with the McNamara approach to government administration, and had the cooperation of Albert Shanker, the head of the United Federation of Teachers. The 1969 contract between the union and the city explicitly notes a joint agreement on professional accountability (Lessinger, 1971, p. 63). The $100,000 contract to design the system was awarded to ETS; reflecting the concerns of the teacher’s union, ETS intended to construct a system which controlled for students’ background characteristics, a point which was criticized by educational psychologist and civil rights activist Kenneth B. Clark. Echoing claims of the ‘soft bigotry of low expectations’ 30 years later, Clark contended that this sort of system would “formalize and give sanction to negative expectations for poor and minority children” (Buder, 1971). Both the New York City Board of Education and the schools Chancellor emphasized that the program would not be used in a punitive fashion against individual teachers or administrators, and that evaluation would be of school performance in the aggregate (Hechinger, 1971). Other accountability designs provoked much resistance among educators. In 1973, the issue of evaluation provoked a six week-long teacher’s strike in Detroit (Stevens, 1973). Shortly after the New York City plan was instituted in 1974, the Schools Chancellor was urged by the president of the Board of Education to address a potential cheating scandal (Buder, 1974). Ultimately, the

97 The contractor in Texarkana, TX – Dorsett Educational Systems Inc. – was found to be instructing students on the same items contained in the evaluation exam (Stake, 1971). In 1970, the Office of Economic Opportunity conducted an 18-site experiment with Performance Contracting. Batelle Laboratories was the first evaluator to release its results, which showed no differences between students taught by contractors and those taught at public schools. The RAND Corporation, which further evaluated the experiment, found no evidence of better achievement at any of the sites (Ascher, 1996).
accountability program was overshadowed by greater concerns surrounding the decentralization of school control and later by the city’s fiscal crisis.

By 1975, accountability was ceasing to be a buzz-word in discussions of educational policy. Schools in large American cities were more actively coping with budget crises as their tax bases migrated to the suburbs. Concerns about educational equity persisted, but were generally expressed in questions of school funding. Three major cases in the California Supreme Court, all titled *Serrano vs. Priest*, were focused on this question; similar cases were argued in many states under headings of ‘funding equity’ and ‘funding adequacy’ (Bollinger, 2003). But perceptions of the concept of school quality were dramatically altered by the application of systems analysis, a change which would set the stage for education reform battles in the 21st century. Because SA, aided by standardized testing and school surveys (each products and producers of previous moral upsurges, as documented above) made possible the direct comparison of schools, the notion of good and bad schools and the distinction between the two became important.

But as school accountability faded into the relative background of educational policy efforts, new technological innovation and political framing was already underway. One educator in 1975 presciently wrote:

“Teachers know all about panaceas. Accountability was the most recent remedy for the ills of American schools. After some obligatory grumbling and retooling, teachers made their peace with it… This interesting notion has useful possibilities. But teachers may not be able to develop them, for a newer panacea is upon them. Productivity has surfaced.”
(Miller F., 1975)

State and federal courts have heard many school funding equity cases in since the 1970s, but there has been one generally constant feature: the expert testimony of Value-Added pioneer Eric Hanushek. He has testified on behalf of the state in twenty such cases, and his research even informed the Supreme Court decision in the 2009 *Horne v. Flores* decision, where the majority opinion suggested, in Hanushek’s language, that school performance measures should focus on output rather than input. This will be revisited in the final chapter.
The next chapter traces the developments of both the technology of evaluation and the attendant political context in terms of teacher productivity. It also leads us to the state of the art with Value-Added Assessment.
Chapter 5: Value-Added in Private: Developing techniques to measure teacher productivity

The last chapter discussed the relationship between educational politics and assessment technology since the earliest days of public schooling. This chapter will resume immediately where the last left off and take us directly through the development of value-added assessment. The focus will be the development of value-added as theoretical model, practical technology and policy tool, as well the discursive and political shifts which accompanied this development.

One of the pioneers of value-added research crystallized the shift in the approach to studying educational systems in a presentation to the American Education Research Association in 2000. Though his reference is the Dallas school district, it could easily be applied to the country as a whole:

   For many years the District and Department followed what was the classical evaluation approach of the late 1960s and early 1970s…Good program evaluation was the key…In the 1980s we tried a comprehensive program to identify and reward effective schools…from 1997 conducting research using teacher effectiveness data (Mendro & Bembry, 2000, pp. 2-3).

Beyond data and technology, how can we understand the last 20 years of education reform from a political/ideological perspective? Anthony Picciano and Joel Spring’s (2013) work on the “Education-Industrial Complex” posits a network of actors which has taken shape in the last 20 years of education policymaking in the US. In seeking to understand how technological changes have been (often uncritically) pushed into American schooling, the authors utilize two important concepts: “flexians” and shadow government. The former term refers to individuals who move
between for-profit, nonprofit, and government employment. Flexians develop networks which take advantage of the executive power of governments, the influence of non-profits, and the economic power of corporations to promote their personal agendas and often to enrich themselves. In an era when governments, particularly at the state and local levels, are in financial trouble and increasingly turning functions over to private firms – what Wedel calls ‘Swiss-cheese government’ (Picciano & Spring, p. 18) – opens many doors for flexians to become part of government networks without being accountable to elected officials or the public. In this way flexian networks – including unaccountable private companies and non-profit foundations – come to constitute a shadow government which can promote its favored ideology.

Picciano and Spring argue that technology is the primary ideology for so-called flexians operating in the education system. However, they are not the only scholars to observe this unbridled enthusiasm for all things digital and algorithmic. Evgeny Morozov (2013) offers a wide ranging critique of such ambitious number-crunchers – many of whom make money working for Silicon Valley firms – who believe that better algorithms and accessible data can solve problems that have challenged the political process for generations. He names this phenomenon ‘technological solutionism’ and suggests that their quest to remake the world in the image of computer software and “the internet” is misguided. Further he argues that if successful, technologically managed politics would undermine the very human freedom that it seeks to enhance. In this chapter, we will see how an exponential growth of computing power coincides with political developments in education and thus enabled value-added to flourish in policy circles.
From Coleman to A Nation At Risk: Economic competitiveness and the narrative of school failure

By the middle of the 1970s, research on schooling had begun to move away from the program evaluation model and systems analysis, even as school systems at all levels continued to rely on these methods. As we have seen, these techniques were deployed partly in reaction to the massive Great Society programs begun under Lyndon Johnson (Moynihan, 1970; Jardini, 2000). Perhaps the most important piece of education research in that era, the 1966 Coleman Report, advanced a controversial position: for student outcomes, in-school factors matter substantially less than out-of-school factors such as students’ home environment and peers. Despite this major finding, the Coleman data’s unique feature of linking students to schools and teachers provided the first opening for research on so-called educational productivity. Among the reactions to that finding lay the seeds of the present debate over teacher quality, though it would take 40 years to appear in its current shape.

Before discussing the details of how assessment technology developed, certain contexts need consideration. First, the Cold War and real or perceived competition with the Soviet Union ebbed as the Vietnam War ended. The end of this ideological conflict may have dampened American policymakers’ appetites for progressive reforms, as they were no longer necessary to maintain capitalism’s appeal among a restive public.99 Thus, even though the envisioned ‘Great Society’

---

99 It could be argued that the decline and eventual demise of the Soviet Union opened up space for the resurgence of neo-Classical economic theories which argue against government interference and in favor of so-called ‘free markets’ in all aspects of social life. Neo-classical resurgence animated many educational reform efforts throughout the 1980s and 1990s, most notably the educational voucher and school choice movements, which can be traced back to Milton Friedman’s writings in the early 1960s. John Chubb and Terry Moe’s (1990) Politics, Markets and America’s Schools popularized the argument for privatization by arguing that more autonomous private schools outperform public schools which are weighed down by the democratic political process.
was never achieved, the pragmatic reasons for pursuing it were gone. At the same time, continued demographic shifts in the nation’s urban centers – “white flight” – accelerated in the wake of attempts to desegregate public schools. These shifts placed great financial strains on many cities, making it unlikely that local leaders, even if they were interested in progressive school policies, would be able to afford them. The fights over school funding in the *Serrano v. Priest* cases in California were partly a result of the fiscal choices of cash-strapped cities.

So public schooling again found itself under duress, but shifts in the nature of the economy also made a dramatic impact on education discourse. The 1970s marked the beginning of the post-Fordist period which was characterized by deindustrialization, decreased reliance on skilled labor in manufacturing, an increase in service sector employment, and technological upgrades to certain occupations. These changes served to increase emphasis on education as an economic growth engine both at the societal and individual level\(^\text{100}\). Human Capital theory, though traceable to the work of Adam Smith, is more commonly associated with economists of the 1960s, particularly Theodore Schultz, Gary Becker and Jacob Mincer. The theory posits education as a form of capital investment; just as upgrades to machinery render production more efficient, so do upgrades to the educational level of the workforce. Formal schooling is offered as the key example of a human capital investment, and is said to offer returns to both individual workers who upgrade their skills and to societal groups that expand access to schooling. Despite substantial challenges to its central tenets (the relation between education and the labor force, and the rational decision-making of individuals), human capital continues to occupy a central

---

\(^{100}\) This link between educational expansion and economic growth was first theorized by Economist Gary Becker in 1962 in his book *Human Capital*. The shift away from manufacturing in the most developed countries makes it easier to argue for this connection, whether or not it actually exists.
place in educational policy-making in the US. Simply put, education is conceived of as an economic imperative both for individuals and groups and for the nation as a whole. This conception has helped fuel panic over schooling since the 1980s.

A related technical innovation was the introduction of a number of international education assessments, which sought to compare the knowledge of students from industrialized countries. The International Association for the Evaluation of Educational Achievement, which has existed in some form since 1958, began attempting to measure and compare educational progress among countries in the early 1980s with its Second International Mathematics (SIMS) and Science (SISS) studies. In 1991 the organization administered its first Reading Literacy Study. The Third International Mathematics and Science Study (TIMSS), conducted in 1995, became the first regular international assessment of educational performance, now repeated every four years, along with the Progress in International Reading Literacy Study (International Association for the Evaluation of Educational Achievement, nd). The Organization for Economic Cooperation and Development (OECD) followed suit in the late 1990s with its own comparative test of all subjects, the Programme for International Student Assessment (PISA). It is important to note that both the impetus for and technology of national assessment originated in the US – the TIMSS program was largely adapted from the National Assessment of Educational Progress (NAEP) which has been administered in the US in some form since the 1960s (Beaton, et al., 2011).

---

101 One motivating force behind the expansion NAEP was the annual release of ACT and SAT test scores. In the 1970s, reaction to media panic over the declining scores on college entrance exams – which tested only those students interested in attending college – crystallized the need for a complete national indicator. The NAEP also underwent significant changes in the 1980s; before 1983, the exam did not allow for comparisons of states or school districts. But technical innovations including Item Response Theory scaling and new sampling techniques allowed for such group comparisons and for tracking of trends in performance (Beaton, et al., 2011).
Concern over educational quality from the mid-1980s forward was first crystallized in a 1983 report commissioned by President Ronald Reagan, *A Nation at Risk: The Imperative for Educational Reform* (ANR). The report framed deficiencies in the US school system in the context of rising international economic competition and stagnating domestic social mobility, and likened existing educational policy to “unilateral disarmament.” The claim that a “rising tide of mediocrity” was threatening America’s economic dominance was based primarily on reported declines in SAT scores and the results of international assessments of academic achievement, which frequently placed the US low among industrialized countries. Given that the federal government does not directly control education policy, ANR urged policymakers in the states to focus on basic competencies and to devote particular attention to science and mathematics education.\(^{102}\)

Notably, the report does not name teachers as the source of the problem, but emphasizes a need to change school policy from the outside – meaning by interest groups not immediately connected to schooling. As during the Cold War, emphasis was on the rigor of coursework, especially in high schools. Other cited issues were the amount of school time spent on studying (as opposed to extracurricular activities) and lowered selectivity among colleges. Teachers are placed among the victims of the school systems and are said to be necessary partners in efforts to change the schools.\(^{103}\) The only problem attributed to teachers directly is a lack of qualifications in needed subject areas such as mathematics, natural sciences, and computer technology. But the

\(^{102}\) Although ANR read more like a manifesto than a policy research report, it was supported by a detailed analysis of students’ high school transcripts, prepared by noted education policy researcher Clifford Adelman.

\(^{103}\) A Nation at Risk is a political document and so one could regard the emphasis on school systems rather than individual teachers as an attempt to prevent a backlash by teachers’ groups. But later reform reports go much further in positioning teachers as an obstacle to the desired changes to the school system.
blame for this is placed at the feet of schools of education, which were charged with emphasizing methodology over content.\textsuperscript{104}

ANR resonated among policymakers and business leaders. Mehta (2013) notes the proliferation of reports by various state education agencies echoing its sentiments. ANR juxtaposed stubborn educational insiders pursuing their narrow interests with external reformers acting on behalf of the public good. The more important shift in language was a contrast to the Coleman Report’s suggestion that schools were unlikely to overcome larger social forces; here, schools were held solely responsible for a perceived decline in American global economic standing and for stagnating social mobility. Similarly, ANR’s emphasis on economic competitiveness and higher level skills required for the jobs of the future raised the stakes of education reform to no less than the future of the country.

Three years later, the focus on the schools in general would be narrowed to educators by another government report. In August of 1986, the National Governor’s Association published and adopted \textit{Time for Results}, which urged state governments to “make it clear to teacher organizations, administrators, and school boards that they must consider new ways of doing business” (Kean, 1986). Like \textit{A Nation at Risk}, it emphasized a connection between education reform, economic competitiveness and social mobility. In a summary of the report’s findings, one author states that the governors’ “highest priorities are jobs and economic development” (Nathan, 1986). But taking the claims of \textit{ANR} as given, \textit{Time for Results} moves to suggest accountability strategies to measure student learning and reward or punish schools for their

\textsuperscript{104} Teacher education was a major focus of many other reports that appeared in the 1980s. Specifically, a number of reports called for a five-year teacher education program and increased efforts to attract and retain mathematics and science teachers (Sadker & Zittleman, 2007).
results. The task force on school management – headed by then governor Bill Clinton – urged school systems to tie certification of principals to student outcomes (Clinton, 1986). Other NGA task forces raised issues of educational technology, school choice, and efficient use of school buildings. Like most educational reform efforts of the 20th century, the outline of the Governors’ action plan stressed educational efficiency – defined as the most progress on test scores and graduation for the least additional cost.

The report’s section on teaching reform stops short of calling for teacher evaluation based on student results, but raises a specter of unionized teachers as an obstacle to reform. The NGA chairman’s overview claims that governors have had to battle with educators to secure real reforms (Alexander L., 1986). In co-chair Thomas Kean’s summary of the teaching task force report, the “total school” is the chosen unit of analysis in order to “foster faculty and staff cooperation.” At the same time, however, the task force repeatedly referred to “organized education groups” and “educational interest groups” whose agendas conflict with “broad public concern.” Specifically, Kean challenges educators to justify calls for increased funding and professional discretion by delivering results, which are only defined in the report by standardized tests. The task force recommends using these measures unless educators can come up with other ways of assessing progress (Kean, 1986, pp. 206-207). Broadly speaking, Time for Results offers teachers and schools autonomy only at the price of accountability.

Not every commissioned report on education advanced the same claims of educational decline, economic peril, and needed accountability. In 1990, the U.S. Department of Energy’s Sandia National Laboratory reported on historical indicators of school performance and projected technological needs of the nation’s workforce. In the first case, the report finds no evidence of a decline in educational quality, though the authors are careful to suggest that improvement was
possible and necessary. SAT score declines were found to be well-explained by a changing population of test-takers, and National Assessment of Educational Progress test scores showed no evidence of decline in mathematics or science (Carson, Huelskamp, & Woodall, 1992, pp. 270-271). On the latter question, the report runs counter to the human capital narrative advanced in *A Nation At Risk* and *Time for Results*. Citing data from a number of surveys, the authors argue that technical skills are not in demand among employers and that, if anything, the production of science and engineering degree-holders is in excess of demand. Further, the authors suggest that even if American students are performing relatively poorly on international standardized tests, their later outcomes such as college attendance, degree attainment, and skilled technical employment are comparatively high. In sum, the *Sandia Report* shows that critics of the American school system at the time were wrong on points of fact. But, as Berliner and Biddle (1995) note, these findings were suppressed by the administration of George H.W. Bush, as they contradicted its education reform agenda.

Critiques of the Coleman Report and the ‘Education Production Function’

In 1966, immediately following the release of *Coleman*, a group of education researchers was convened to discuss the survey and its findings. These meetings, called the Harvard Seminar on the Equality of Educational Opportunity Report (SEEOR), sought to reanalyze the vast data collected and further interrogate the main findings. In general, these seminars confirmed what the original report had said about the relationship between school inputs and outputs, that variation in academic performance was higher *within* schools than *between* schools and thus more attributable to student characteristics than to anything that could be easily changes about school resources, including staff.
But one group of critics suggested that the report’s data on school quality and its analytic methods were inadequate for discovering the relationship between school quality and student academic achievement. They further argued that the data and methods systematically favored the ‘no school effects’ conclusion. Among the problems, these critics suggested that teacher characteristics should have been included in the group of school input variables (Hanushek & Kain, 1972). Eric Hanushek, at the time a graduate student at MIT, was one of the participants arguing this position.¹⁰⁵ His doctoral dissertation was his first attempt to revise the Coleman analysis. Therein, he further developed the ‘teacher effect’ concept in regards to racial differences in academic achievement. Shortly thereafter, building on an argument which couldn’t be directly tested with Coleman Report data, Hanushek used longitudinal data to estimate teacher effects (Hanushek E., 1970). In a subsequent article, he used achievement gains between the first and third grades as an outcome measure (Hanushek E., 1971). This change in approach marked the birth of value-added assessment in education.¹⁰⁶

Other researchers saw problems with the measurement of educational productivity similar to those of the systems analysis approach. In a 1970 paper, Samuel Bowles discussed the theoretical and practical issues with attempting to quantify what and how well schools produce. At the conceptual level, researchers lack a clear theory of student learning with which to specify the production process. Similarly, even if learning outcomes are accepted as one goal of schooling,

¹⁰⁵ Hanushek also figures into the last part of the story in the previous chapter. He was a consultant for the RAND corporation from 1969-1973, during which time he published a report titled “The Value of Teachers in Teaching.” In his expert witness testimony in Vergara v. California, examined in the next chapter, we learn how this research figured into his first expert witness testimony in the 1973 Serrano v. Priest case (Mar 24 AM, 0h18m).

¹⁰⁶ Even at this early point, the policy implications of Value-Added were quite clear. The author suggests that swapping teachers at the bottom of the verbal ability scale for those at the top would produce dramatic achievement gains (Hanushek E., 1971, p. 286).
there are certainly others. Bowles called school output “multidimensional with a vengeance” (Bowles, 1970, p. 23). Practically, researchers could not control variation in the school situation or the schooling process. Further, schools do not behave in the ways efficiency-driven researchers would like them to, frequently making decisions based on criteria other than the maximization of output. Bowles argued that available data are not precise enough to support even an approximate model of student learning. Beyond this, he follows Coleman and others in stressing that given how much learning takes place outside of school, the potential productivity of schools is likely to be overstated by those interested in interrogating the concept of educational production (pp. 40-41).

But to move beyond simulations and secondary data analysis, Value-Added assessment would require massive data collection projects. The city of Dallas and the state of Tennessee were the first testing grounds for such efforts.

Early Value-Added Systems: DVAAS and TVAAS

Dallas, Texas was the first municipality to evaluate education using value-added techniques (Stone, 2000). Beginning in the late 1960s, and accelerating after a lawsuit in 1971, the Dallas Independent School District was charged with assessing the effectiveness of its desegregation initiatives. As was shown in the previous chapter, accountability was a policy buzzword at the time, and was enmeshed with desegregation efforts. William Webster was in charge of research and development for the Dallas schools and was responsible for compliance with state-mandated accountability (Cunningham, 1997). He is regarded as primarily responsible for the Dallas Value-Added Accountability System (DVAAS).
In 1984, the district began using a version of value-added based on regression techniques to evaluate schools.\textsuperscript{107} That system was abandoned and replaced in 1990 by another which used hierarchical linear modeling; another iteration of DVAAS, which could be used to evaluate individual teachers, was field tested in 1996.\textsuperscript{108} Despite the ability to assess teachers, the Commission which created the system emphasized identifying effective schools and merit pay (Webster & Mendro, 1997).

By this time, many researchers had expressed concern about the use of student achievement data in evaluation of teacher performance, owing to the fallout over merit pay policies in previous decades (Webster & Mendro, 1995, p. 393). Technically proficient auditors expressed reservations about the Dallas system, noting that its estimates of teacher effects would be unreliable (Thum & Bryk, 1997). More holistic critics suggested potential unintended consequences of an evaluation system based on high-stakes multiple choice tests – including widespread cheating, narrowing of the curriculum, and negative impacts on school culture (Sykes, 1997). The architects of the system responded that despite its limitations, the value-
added assessment system would be a fairer means of assessing schools and teachers than other systems currently in use.\textsuperscript{109}

Dallas’ foray into Value-Added assessment is important not only for its novelty. The model of accountability created there would substantially influence the Texas model implemented by George W. Bush, which would then be replicated in NCLB. Sandy Kress, a Democratic Party operative in the 1970s and 1980s, was president of the board of trustees for the Dallas Public Schools and a zealous believer in the power of accountability (Donald, 2000). Before becoming its president, he led the board’s Commission for Educational Excellence which recommended Mendro’s Value-Added system and gave it consequences for schools and teachers (Bembry, Bearden, & Mendro, 1997). After four divisive years serving on the school board, Kress resigned, and quickly became engaged in promoting educational accountability across Texas. He earned the favor of George W. Bush, and despite partisan disagreements, was the chief promoter of the “Texas miracle” – overall test score gains and a closing of the racial achievement gap.

\textsuperscript{109} Representative of the status of ‘rigorous science’ in educational research discussed in the next section, Webster and his colleagues devote careful attention to the technical criticisms of Thum and Bryk, while regarding Sykes’ thoughtful and perhaps more relevant critique of the Dallas Value-Added System as ‘Pie in the Sky.’
attributed to the Texas accountability system\textsuperscript{110} – during Bush’s presidential campaign and became the key architect of NCLB’s accountability provisions.\textsuperscript{111}

The first state to implement data collection for value-added assessment was Tennessee, which has never been shy about novel education policy. In 1983, the state enacted a merit pay system called the Career Ladder System. Two years later, it hosted perhaps the largest randomized educational experiment in US history, Project STAR, to assess the effects of class size on student achievement. Ceperly and Reel (1997) note that Tennessee was pushed into accountability reform by a 1988 lawsuit over school funding equity\textsuperscript{112}. As a result of the lawsuit, legislators needed business support to raise taxes and equalize school funding. In return, they promised the state’s business roundtable that schools and teachers would be held accountable for performance.

William L. Sanders of the University of Tennessee\textsuperscript{113} had begun pilot studies of a value-added assessment system as early as 1982 with funding from the state’s department of education (Ball, 2013).

\textsuperscript{110} The “Texas miracle” is now widely regarded as a myth and a case study in how testing regimes can pervert the educational process. Many point to Texas as the archetype of “educational triage” – focusing instruction on students who are on the verge of being ranked proficient by a test – while ignoring both those who will definitely pass and definitely fail (Booher-Jennings, 2005). The parallel in studies of social service accountability is “creaming” – prioritizing the easiest cases when one is rated on the number of successful outcomes (Miller, Roby, & van Steenwijk, 1970; Cohen N. A., 2014). Others note that Texas saw a dramatic increase in the number of special education placements as average scores were on the rise – special education students’ scores did not count in accountability metrics (Haney, 2000). Observers likened this practice to a form of creative accounting which bolstered Texas’ reputation without improving its educational quality (Bernstein J., 2002).

\textsuperscript{111} Illustrative of the Picciano and Spring’s ‘flexians,’ Kress now works as a lobbyist for a number of private education firms including Pearson, Teach for America, and Wireless Generation (Vasquez-Heilig, 2013).

\textsuperscript{112} The lawsuit, Tennessee Small Schools v. McWherter, went through a number of appeals which concluded in 2006.

\textsuperscript{113} Before his research moved to education, Sanders was an agricultural statistician in Tennessee whose research focused on reproduction in livestock, specifically how to parse out the influence of genetics from that of environment. In the late 1970s, then Governor Lamar Alexander wanted to use student achievement data for the state’s merit pay program. Amid controversy over the
The result of these efforts was the Tennessee Value-Added Assessment System (TVAAS). The system was immediately popular with state lawmakers, and Sanders became a popular lecturer both inside the state and elsewhere. The primary argument for its superiority over other means of test-based evaluation was that it used gain-scores or student achievement growth, rather than raw test score data, to compare school effectiveness. As in Dallas, Tennessee regarded this as fairer because it seemingly levels the playing field between schools by not punishing those that serve low-scoring students.

When Tennessee passed its Education Improvement Act in 1991, Sanders’ system was the centerpiece of the legislation. The law specifically mandated the use of TVAAS. The system made use of norm-referenced tests the state’s Comprehensive Assessment Plan (TCAP) for its annual data. Between 1991 and 2003, the state paid over $5 million to Sanders’ research outfit just to cover the data analysis contract (Smith, 2004). The first few years of data collection still focused on school-level analysis, but in 1994 student and teacher information were linked in the database, allowing for Value-Added measurement of individual teachers. Since 1993, value-added reports on schools and school systems have been part of the public record in Tennessee; teacher reports only commenced in 1996 and are only distributed to teachers and administrators.

Appropriateness of using such data, Sanders used agricultural value-added methods with data from Knox County to make the case that the technique was viable. The initial study, concluded in 1984, failed to gain the attention of legislators (Archer, 1999). A departure from earlier education evaluation laws, the Tennessee EIA specified that TVAAS would be used to evaluate individual teachers’ contributions to student achievement. TVAAS’ use of norm-referenced tests is also a significant detail. Norm-referenced tests are not designed to reflect mastery of content areas; they instead seek to rank students by using small batches of questions covering broad areas of knowledge. The SAT is an example of a norm-references test. Tennessee’s norm-referenced tests compared Tennessee students with other students around the country. Criterion-referenced tests are designed to assess mastery of specified curriculum goals. Tennessee’s criterion-referenced exams were aligned with its state curriculum (Baker & Xu, 1995, p. 3). Thus, the tests used to assess teacher quality were not actually based on what teachers were expected to teach.
Furthermore, the state placed limitations on which student achievement scores could be used to determine teachers’ value-added scores (Baker & Xu, 1995).

Sanders’ team collected annual data on all of the state’s public school students. Sanders recognized that in addition to the evaluation function the data would serve within Tennessee, the scope and nature of the data would have effects on educational research well beyond state borders. Many of the conclusions drawn from early analysis of Tennessee data have become recurring themes regarding the merits of Value-Added methods, most notably the claim of residual effects of ‘highly effective’ teachers in subsequent years (Sanders & Horn, 1998). But at this stage, Sanders was still careful to note that while VAA data could be valuable for understanding how teachers impact student learning, it should not be used to make high-stakes decisions:

“We hope the data are used for diagnostic purposes…This is about trying to improve our schools, not about embarrassing teachers (Hill D., 2000).”

“To my knowledge, no teacher has been fired where the reports were the justification for it. This is not about firing people. This is about measurement, about producing what I call the river of diagnostic information to show individual teachers where their relative strengths and weaknesses are (Clowes, 1999).”

Echoing the wisdom of Coleman, Sanders also remarked that schools and teachers should not be held to account for societal problems beyond their control (Archer, 1999).

Used for consequence or not, TVAAS was subject to criticism early on because it did not account for student background variables in the determination of teacher effects, and for its questionable handling of missing data (Darlington, 1997). The method used in the early Dallas system did include background measures, similar to systems currently in place in New York and Los Angeles. But Sanders devoted much time to arguing that background characteristics were superfluous to a model which included individual students’ prior test scores (Sanders & Horn,
Another important criticism focused on the complexity of the statistical techniques, which rendered the assessment incomprehensible to nearly everyone directly affected by it (Baker & Xu, 1995; Popham, 1997). The TVAAS system provided an important model for this method of teacher and school evaluation, but subsequent political developments would push Value-Added into the education reform spotlight.

Clinton, Bush, and NCLB: Education Science and the Focus on Teachers

While much criticism is leveled at No Child Left Behind (NCLB), the educational program enacted by President George W. Bush, its emphasis on measurable outputs and accountability has its root in the educational agenda of President Bill Clinton. The Improving America’s Schools Act (IASA), Clinton’s reauthorization of the Elementary and Secondary Education Act, for the first time tied Title I funding to schools’ adoption of testing programs, and required districts to take corrective action against schools not meeting improvement standards (Education Week, 1994; Thomas & Brady, 2005). Clinton’s ‘Goals 2000’ legislation, signed just before the IASA, included a grant program intended to leverage state-level cooperation with the president’s

---

115 The general consensus among VA researchers is that the models should include controls for background characteristics. I’ve already shown in chapter 2 how background characteristics at the school level impact teacher VA rankings.

116 Some trace this lineage back even further to the convening of the Charlottesville Education Summit by President George H. W. Bush in 1989. But the goals set by this summit were only given any legal force after Clinton, himself a key architect of the summit, was elected in 1992. Clinton had also been influenced by his experience in the National Governor’s association – during which he was a contributor to the above-mentioned Time for Results. Bush’s legislative attempt, America 2000, failed due to opposition from conservatives in his party, who argued it was an incursion on states’ legislative territory. Nonetheless, the bill was notable for its emphasis on uniform national academic standards (McDonnell, 2005).
education agenda (Heise, 1994) – bearing much resemblance President Barack Obama’s Race to the Top program 15 years later.\textsuperscript{117}

But the overlap among the education policies of two Republicans and one Democrat itself needs explanation. Goals 2000’s emphasis on standards-based reforms was mirrored in No Child Left Behind, as was its emphasis on the expansion of educational technology. In their similarities we can see the long shadow of A Nation at Risk, both Bushes and Clinton made use of the rhetoric of an underprepared workforce in need of experience with new technology. Looking at the evolving role of the federal government in education, McDonnell (2005) suggests that the common experiences of Lamar Alexander (the elder Bush’s secretary of education), Bill Clinton, and George W. Bush as governors in southern states helps explain their similar attraction to standardized testing as a means of measuring academic progress. As we have already seen, two of the three southern states (Texas and Tennessee) gave birth to Value-Added Assessment. Rudalevige (2003) contends that in Congress, Republicans and many Democrats had grown frustrated with the lack of results – however defined – from money spent on education. In her analysis of the same period, Elizabeth DeBray (2006) notes that in the 1990s, conservative think-tanks began to take clear positions with respect to education policy – specifically promoting school choice, vouchers, and other market-based reforms. Organizations like the American Enterprise Institute, the Thomas B. Fordham Foundation, and the Hoover Institution began drafting education policy platforms, which until then had been the province of more centrist and

\textsuperscript{117} After the election of a Republican majority in the House of Representatives under Speaker Newt Gingrich, congressional Republicans fought Goals 2000 – particularly its emphasis on national education standards.
liberal groups (pp. 24-26).\textsuperscript{118} This active conservative influence drove education policy further to the ideological right and fostered a hostility toward both public schools and unionized teachers.

This hostile climate took one form in the push for scientifically based research in education. The first attempt in this direction came with the 1999 Reading Excellence Act. Another competitive grant program for the states, the REA defined ‘scientific reading research’ as “grounded in rigorous, systematic and objective procedures…employ[ing] systematic, empirical methods that draw on observation or experiment, involve rigorous data analyses that are designed to test the stated hypotheses and justify the general conclusions drawn” (Baez & Boyles, 2009, p. 6).

Another piece of legislation, drafted in 2000, aimed to reform the Department of Education’s Office of Educational Research and Improvement (now the Institute for Education Sciences), included a similar definition of scientific research, but in the context of a bill which would affect the practice of research (Eisenhart & Towne, 2003).\textsuperscript{119} But these were precursors to a much more wide-reaching debate over what constitutes scientific research in education.

In November 2001, the National Research Council released a report titled \textit{Scientific Research in Education}. In its treatment of what constitutes rigorous research, the council was carefully ecumenical on questions of methodology – making sure to note that non-experimental and qualitative methods could be deemed rigorous. Yet certain inconsistencies are apparent. At the outset of the report, they argue that science can be value-free and insulated from political and

\textsuperscript{118} Before the 1990s, conservatives espoused Ronald Reagan’s position on federal education policy – the abolition of the Department of Education and the total handoff of schooling to the states. Clinton’s re-election in 1996 further confirmed to many in the Republican party that such an anti-education agenda was not palatable to the American electorate.

\textsuperscript{119} The “Castle Bill” also contained standards for qualitative research methods, but which proscribe them to only those questions are not amenable to quantitative analysis (Eisenhart & Towne, Contestation and Change in National Policy on ‘Scientifically Based’ Education Research, 2003, p. 32).
sociological forces, yet they use the history of standardized testing (discussed at length in the previous chapter) as an example of the steady – and apolitical – accumulation of scientific knowledge in education (National Research Council, 2002). Perhaps more significantly, the examples of rigorous research offered in the report are overwhelmingly quantitative and experimental. Furthermore, the council report emphasizes that scientific research ultimately aims at establishing cause-and-effect, implying (and devoting considerable attention to) the sort of experimental and quantitative methods that have been ascendant in public policy since the 1970s (p.101; p.110-126).

But it was two successive pieces of legislation, No Child Left Behind Act of 2001 and the Education Sciences Reform Act of 2002, which unequivocally enshrined experimental and quantitative methods as the standard for educational research. Scientifically based research is defined in NCLB as “using empirical methods that draw on observation or experiment” and “evaluated using experimental or quasi-experimental designs…with a preference for random-assignment experiments, or other designs to the extent that those designs contain within-condition or cross-condition controls (NCLB 2001, 2002).” The full text of the law refers to scientifically based research over 100 times, mostly in the sections concerning disadvantaged students and their schools, and those outlining “teacher quality” (Manna & Petrilli, 2008). What this meant in practice was that schools and districts would be obliged to choose services and programs which were vetted using experimental and quasi-experimental design research if they were to receive funding from the federal government under the Elementary and Secondary 

---

120 Notably, the sections of the law which pertain to expansion of charter schools and educational technology contain almost no references to scientifically-based research (Manna & Petrilli, 2008, pp. 67-68). Apparently, the standard of experimental design was not meant to be equally applied to the favored projects of the Bush administration. This provides additional evidence of the influence of Picciano and Spring’s (2013) educational flexians.
Education Act. Thus, indirectly, education researchers who wanted to influence educational practice would have to meet this standard (Barhouse-Walters, Lareau, & Ranis, 2009a).

The Education Sciences Reform Act (ESRA) went further in establishing what kinds of research the federal government would fund based on similar criteria to those of NCLB. The act abolished the Office of Educational Research and Improvement and replaced it with the Institute of Education Sciences (IES). The force of the law for education research was that the Department of Education would only fund projects which adhered to its definition of scientifically valid research. While the ESRA definition is slightly broader than NCLB, it shows a clear preference for experimental and quasi-experimental designs and emphasis on questions of causality and program evaluation (Barnhouse-Walters, Lareau, & Ranis, Appendix B: The Definitions of "Scientifically-Based Research" in the Education Sciences Reform Act of 2002, 2009b). A related development was the What Works Clearinghouse, an initiative of the IES which evaluates research on educational interventions based on strict methodological criteria – meaning only randomized controlled trials and quasi-experimental quantitative studies – with the intention of helping “students and schools meet high standards…by providing educators with credible and reliable evidence…to make informed decisions” (Institute of Education Sciences, nd). The WWC was based on an earlier effort by Chester Finn – Ronald Reagan’s Assistant Secretary of Education – to strengthen the practical utility of education research. Walters (2009) suggests that these successive redefinitions of scientific research are a culmination in the effort to reign in certain tendencies which had gained currency among education researchers since the 1960s –

---

121 Perhaps reflecting President Bush’s intentions for the IES, he chose Grover “Russ” Whitehurst, an experimental child psychologist, as founding head of the institute. Whitehurst is a fellow at the conservative Hoover Institute, and until recently was director of the Brookings Institution’s Brown Center for Education Policy. During his tenure in 2012, he terminated Diane Ravitch as an unpaid senior fellow.
postmodernism, critical theory and interpretivism. In 2004, IES also provided $760,000 in startup money to found the Society for Research on Educational Effectiveness (SREE), an organization committed to “applying the procedural norms of science to the study of educational problems” and focused on the “cause-and-effect relations important for education” (Society for Research on Educational Effectiveness, nd).

What makes such an effort hostile toward educators is its emphasis on distinguishing legitimate “scientific” knowledge from other sorts of discourse about education. When education is turned into a problem to be solved by scientists, it implies that those presently responsible for the practice of education are not equipped to the task – or are not producing the desired results. Thus, while some randomized field trials may produce valuable insights about how to reform educational practice:

“one cannot ignore the politics associated with constituting knowledge as scientific, particularly the boundaries created around who and what can be said to constitute such knowledge and the differentiating status such boundaries ensure. For monopolizing knowledge, any knowledge, but particularly scientific knowledge, restricts opportunities for others to think, in effect ensuring a large class of “unskilled” workers (which is all that teachers and other mere practitioners will become)…” (Baez & Boyles, 2009, p. 30).

In this way, the attempt to define education policy as an “evidence-based field” can be seen as an attempt to wrest control away from educators themselves and into the hands of a group of analysts operating in a particular paradigm.

Perhaps unsurprisingly, this portion of NCLB easily passed through a normally contentious Congress. Legislators have long been enamored with social engineering strategies – turning complex processes into simple indicators of causality and efficiency. James Scott observes the phenomenon in his classic work Seeing Like a State (Scott, Seeing Like A State: How Certain
Schemes to Improve the Human Condition Have Failed, 1998). Looking specifically at American schooling, Wise (1979) notes the tendency toward and limitations of administrative solutions to educational problems. But beyond this push for ‘education science,’ NCLB directly targeted educators through its emphasis on teacher quality. Indeed, ‘highly qualified teachers’ are the sole focus of Title II of the law, and the phrase features prominently throughout the entire document.

Manna and Petrilli (2008) suggest that three think tanks were significant in crafting this focus: Education Trust, the Fordham Foundation, and the National Commission on Teaching and America’s Future (NCTAF). Each provided summaries of research on teacher quality during legislative hearings prior to the passage of NCLB. EdTrust and Fordham focused their summaries on research which favored Value-Added Assessment as a means of measuring teacher quality. EdTrust relied heavily on Sanders’ work on the long-term impacts of effective teachers, featuring a chart from his 1996 study on the cover of their research summary. They also cite research by Erik Hanushek, the early proponent of education production function research discussed above. The conservative Fordham Foundation - then headed by Chester Finn – cited many of the same sources as EdTrust, including Sanders and Hanushek, in order to emphasize the connection between teacher quality and student achievement. What distinguishes the two is that Fordham Foundation cited other research which advocated deregulation strategies to address

---

122 The ‘highly qualified teacher’ provision specified that by 2005-06, all teachers covered under the ESEA would be required to possess a Bachelor’s degree and full certification in the state where they were teaching, in addition to meeting requirements for subject matter expertise. Notably, the provision did not apply to charter schools, and a 2002 modification to the law exempted trainees in alternative certification programs – such as Teach for America – from the requirement (NCLB 2001, 2002; Zeichner, 2013).

123 The Sanders paper cited in the EdTrust and Fordham research summaries (Sanders & Rivers, 1996) makes an argument very similar to that of another high profile study done in 2013 by Harvard economist Raj Chetty, discussed below.
the teacher quality problem. NCTAF, directed by Linda Darling-Hammond at the time, focused primarily on upgrading teacher credentials and improving teacher preparation programs, an approach which the Fordham Foundation directly attacked (Manna & Petrilli, 2008, pp. 77-80).

Ultimately however, VAA methods were not mandated in the accountability provisions of NCLB, likely due to the influence of congressional Democrats (perennially reliant upon the support of Teachers’ unions), who were crucial in its passage. George W. Bush had to break with the right wing of his party, which opposed the law’s mandated federal testing requirements. Nonetheless, much like the Interstate Commerce Commission hearings had done for Scientific Management, the attention of federal hearings gave Value-Added Assessment currency among education reformers and policymakers. Perhaps more importantly, NCLB’s impossible mandate for “100 percent proficiency” for all students and its requirement for annual testing in multiple subjects gave the Obama administration leverage to push outcomes-based teacher assessment on the states. Further, according to the expert testimony of Harvard economist Thomas Kane in the 2014 Vergara v. California case – discussed in detail in the next chapter – this testing requirement also created the volumes of data value-added researchers needed to study year-to-year achievement gains. This leads to the early research promoting value-added as a policy tool.

The Widget Effect, NCTQ, and Political Research on Value-Added

As noted above, the annual testing mandates set out by No Child Left Behind facilitated a flood of new student test score data like those produced in The Coleman Report. Researchers interested in Value-Added assessment could begin to analyze state testing data, which could now generate

124 DeBray (2006) notes, for example, the key role played by Democratic Senator Ted Kennedy in the passage of NCLB. Connecting the history of accountability policy in Texas, Sandy Kress played a key role in persuading Kennedy to support NCLB (Stanford, 2013).
the gain-scores for the models. As one might expect, one of the first major post-NCLB value-added studies was conducted by Eric Hanushek.

Hanushek and colleagues relied on a large set of data (over 400,000 students across two cohorts in grades 3 through 7) from the Texas schools system (Hanushek, Rivkin, & Kain, 2005). From the outset, the paper is conceived as a counter to education research since Coleman, and specifically to the enthusiasm for class-size reduction policies since the 1983 Tennessee STAR experiment. They make a bold claim for value-added, stating that because it utilizes individual student test scores in multiple years, the models “are able to identify the impacts of schools and teachers uncontaminated by the many unobserved family and other influences that have plagued past research” (p. 419). The authors do note that the “central estimation problem” with VA methods is that the processes of assigning teachers and students to schools and classrooms are not random and thus subject to bias (p. 423). This caveat anticipates criticisms offered by Jesse Rothstein (2009) and others which have already been discussed in chapter 1 above. A further limitation of the study is that students were not linked to individual teachers, thus the analysis relies on differences between grade-levels to infer teacher quality differences.125

Hanushek and his colleagues use teacher turnover events to demonstrate the existence of teacher effects on student achievement.126 Based on their analysis incidents of teacher turnover, they

---

125 Grade-level value-added is calculated by the difference between two cohorts test scores in the same grade. Grades (and by extension, teachers) are then compared in terms of grade-average value-added scores. Converting these grade-average value-added scores into a measure of the variability requires a number of very strict assumptions, including that the variation in teacher quality is the same for all grades and cohorts and that there are no strong complementarities between specific students and teachers (Hanushek, Rivkin, & Kain, 2005, pp. 426-427).

126 Teacher turnover – incidents when a teacher moves from one school or grade to another – is exploited by Hanushek (and subsequent VA researchers) because in theory it provides an opportunity to show how changes in teacher quality correlate with changes in student achievement. If we assume that nothing else changes about the school, the students in it, or the
report that a one standard-deviation increase in teacher VA in a school grade results in a .11 standard deviation growth in student achievement in reading and a .095 SD achievement growth in math. Subsequent analysis in the paper show that even these modest findings only apply to elementary school grades, and more consistently in math. Nonetheless, they contend that these are lower bound estimates of teacher effects and that “thus there can be little doubt that teacher quality is an important determinant of mathematics and reading achievement” (2005, p. 440). They further conclude by arguing that policy emphasis on teacher credentialing and class size reduction are misguided given their estimates of teacher quality effects exceed observed impacts of credentials or class size.

Another important post-NCLB value-added study picks up on the conclusions advanced by Hanushek et al. In *Identifying Effective Teachers Using Performance on the Job*, Robert Gordon, Thomas Kane and Doug Staiger (2006) begin their study of teacher effectiveness by positing:

> “ultimately, the success of US public education depends upon the skills of the 3.1 million teachers managing classrooms in elementary and secondary schools around the country. Everything else – education standards, testing, class size, greater accountability – is background…Without the right people standing in front of the classroom, school reform is a futile exercise” (p. 5).

They propose that the evidence linking teacher credentials to student achievement is weak, and that policymakers need to require that teachers demonstrate minimum competency on the job. They define competency as student achievement growth and measure it using a value-added model which controls for students prior test scores and demographics (similar to the DVAAS teacher herself when she arrives in her new position; it is possible to regard the increase or decrease in student performance as evidence of a change in teacher quality attributable to the new teacher. This approach is refined and utilized by economist Raj Chetty and his colleagues in a paper discussed in detail in the next chapter.
model above). Their study relies on data from the Los Angeles Unified School District from 2000 to 2003. Although smaller than Hanushek et al.’s Texas data (only 150,000 students across four cohorts) the LAUSD data are in student-teacher linked format which Hanushek cited as a critical component. Yet, as opposed to Hanushek et al., who rely on numerous tables and model specifications, Gordon et al. presents only two figures. The first plots the distribution of teacher impacts on average student achievement by teacher credential level; the second plots teacher impacts by VA effectiveness quartile. The first figure shows largely overlapping distributions, suggesting that credentials are not related to teacher effects on test scores. The second shows four relatively distinct distributions, suggesting that prior measures of teacher impacts on student test score gains identify teachers who increase student achievement.

The presentation of evidence concludes on page 9 of the 30-page report. The majority of the paper focuses not on whether value-added identifies effective teachers, but on how VA could be leveraged as a policy for improving US public education. The first prescription they offer is for states to ‘reduce barriers to entering teaching’ by broadening the NCLB definition of high-quality teachers and specifically through the use of alternative routes to certification. They buttress this point by pointing to the Teach for America program and by raising alarm of a

---

127 In their recommendations for improving teacher evaluations, they strongly advise that VA models control for background characteristics.
128 Another figure is presented much later in the document, which plots teacher impacts by years of experience (p. 28). That is used to justify the two-year probationary period, but also suggests the negative impact of constantly replacing two-year veterans with brand new teachers.
129 This is an early incidence of the all-too-common circular reasoning employed by VA proponents. If teachers impact gains in student test scores in prior years, chances are they will do so again in subsequent years. Even if one accepts that this is an unbiased measure, the question remains whether boosting test scores wholly constitutes teacher effectiveness.
130 Teach for America is a two-year teaching program which has not been shown to generate a significant number of career educators. A recent rigorous study commissioned by Teach for America found that 87 percent of TFA recruits trained as teachers did not plan to remain in teaching (Clark, Isenberg, Liu, Makowsky, & Zukiewicz, 2015).
‘coming teacher shortage,’ suggesting that baby-boom generation teachers will retire in droves and that interest in teaching is declining (Gordon, Kane, & Staiger, Identifying Effective Teachers Using Performance on the Job, 2006, pp. 10-12).\textsuperscript{131}

The second recommendation is that schools make it harder for the least effective teachers to be granted tenure. Specifically, they recommend that tenure be denied to teachers in the bottom VA effectiveness quartile after two years. Foreshadowing later research on the long-term impacts of effective teachers, the authors speculate that such a policy would lead to cumulative test score increases and lifetime earnings gains of up to $169,000 per high school graduate and up to $507 billion dollars per year in nationwide earnings gains! (p. 15) Their third recommendation focuses on achievement gaps and what they refer to as maldistribution of effective teachers. They suggest that high-VA teachers be offered substantial pay increases for teaching in low-performing schools. Since there is no conclusive research on the motivating effect of pay increases among teachers, the authors suggest that top quartile teachers be given at least $15,000 per year if they are willing to teach in high-poverty schools.

Their final two recommendations are more specific to teacher evaluation and more directly address the use of VA. The first of these is that teacher effectiveness needs to be defined by multiple measures. Here, the authors argue that, while imperfect, VA rankings based on standardized tests constitute the most important and only “objective” measure of effectiveness (p. 18). They propose that other measures – classroom observations and reviews of student work

\textsuperscript{131} As of this writing, there is little evidence that the “coming teacher shortage” ever arrived. Brenneman (2015) suggests that while there are subject area shortages in Math, Science, bilingual and Special Education fields, there is no national crisis in teacher supply. Shortages in Math and Science have causes well beyond teacher certification, and so changes those processes would be unlikely to resolve these specific problems.
– could supplement assessment of teachers, but that these subjective measures are of secondary importance. In their opinion, VA or some other measure based on student test scores should constitute the preponderant factor in teacher evaluation (p. 19). To facilitate evaluation of all teachers, they recommend that standards and related assessments be developed for all subjects in all grades so that testing can happen every year. Finally, they recommend that schools be provided funding to build the data infrastructure necessary to evaluate teachers based on student test score growth (pp. 22-23). Though Identifying Effective Teachers does not rely on much value-added data, it is politically far more important. This is perhaps the first incidence of proposing VA as a policy solution to perceived educational shortfalls. 132 In its use of dramatic statistics ($507 billion dollars in national income) and dire warnings (The Coming Teacher Shortage), this report cast value-added as the cornerstone of a new strategy to transform American education. It also gained one of its authors, Thomas Kane, the attention of the Bill and Melinda Gates Foundation. This relationship led to the largest educational policy experiment in American history – the Measures of Effective Teaching Project.

But another subsequent study, although not about Value-Added specifically133, dropped a further bombshell on the question of teacher effectiveness. The Widget Effect was published in 2009 by an education policy advocacy group called The New Teacher Project (TNTP)134. Like Identifying Effective Teachers, Widget begins with an ominous quote (from 1936):

---

132 Further evidence of its intended policy impact, the report estimates the costs of implementing its recommendations ($3 billion per year), and provides answers to common criticisms from teachers’ organizations. (pp. 24-29).
133 When The Widget Effect refers to the consequences of ineffective teaching, it utilizes VA studies conducted by William Sanders and Eric Hanushek.
134 TNTP was founded in 1997 by former Washington DC schools Chancellor and notable education reformer Michelle Rhee. The Widget Effect research was supported by funding from the Gates and Walton Foundations, among others.
“There are at least ‘several hundred’ incompetents now in the school system [says the superintendent]. Other observers think there are several thousands, while still others insist that ‘several’ would be nearer the mark. Whether these incompetents were unfit to teach at any time, or have been rendered unfit by the passing years, is a matter of opinion. The question is, why are they allowed to remain?” (Weisberg, Sexton, Mulhern, & Keeling, 2009)

This opening places The Widget Effect in a long tradition of doom-saying educational reports beginning with Laggards in Our Schools over a century earlier. Sounding much like A Nation at Risk, it characterizes the failure to identify and act on teacher effectiveness as a “national failure.” It defines the ‘widget effect’ as “institutional indifference to variations in teacher performance” and characterizes this as both “deeply disrespectful to teachers” and as a “gamble with the lives of students.” (pp. 4-6) The report relies on analysis of district teacher evaluation systems and on survey data from about 16,000 teachers and administrators in 12 school districts across four states (AR, CO, IL, OH). Based on the survey results, the authors point to five indications of the Widget Effect operating in schools: 1) nearly all teachers receive “satisfactory” or better ratings, 2) not enough is done to identify and reward the best teachers, 3) not enough feedback is given for teachers to improve, 4) new teachers are not given enough professional support, and 5) ineffective teaching goes unaddressed. It seems from the introduction of the

135 The story, from The New York Times, is far more banal than the quote suggests; it provides a history and description of the New York City teacher tenure system (as of 1936) and contains one (unfounded) suggestion from an assistant superintendent that there are “several hundred” incompetent teachers out of 36,000 in the NYC school system. The source of incompetence suggested in the story is that many teachers are on the job for too long; the article cites a recommendation that the Teacher Retirement Board takes steps to retire more teachers to make way for new recruits (Bernstein V., 1936).

136 Only ten of the twelve surveyed districts provided teacher evaluation system data which buttress the study’s major point – that most teachers are ranked satisfactory or better. DiCarlo (2014) points out that the surveys on which Widget is based were voluntary and that in the districts surveyed there were about 41,000 active teachers; about 15,000 teachers responded. This represents a 35% non-random response rate. A review of the study has also questioned the selection of school districts, arguing that they do not represent the full scope of approached to teacher evaluation (Pechone & Wei, 2009).
report that this last point is the most important. The authors seem to set as their aim the removal of so-called incompetent teachers – even if couching this aim in a historical reference.\(^{137}\)

*The Widget Effect*’s first point is simple. Analysis of existing teacher evaluation systems shows that most teachers are rated satisfactory or better.\(^{138}\) In two of the districts studied, this holds true even when the schools themselves are faring poorly (pp. 11-12). However, the other points are not as solid. They are based on voluntary responses to an online survey. Further, many of the statistics refer to subpopulations of the 15,000 teachers or 1,300 administrators, making the voluntary survey results even seem even more dubious.\(^{139}\) Taken at face value, the surveys suggest that a substantial group of teachers and administrators feel teacher evaluation systems do not adequately capture the differences among teachers in terms of effectiveness. More importantly, a substantial majority of surveyed teachers did not receive feedback on how they

\(^{137}\) A similar tone is struck in another highly politicized piece of media. The documentary film *Waiting for Superman*, which is at heart an advocacy piece for the charter school industry, takes as one of its foci the inability of many US public school systems to fire ineffective teachers. The filmmakers describe what they term the “dance of the lemons,” in which tenured teachers deemed ineffective (though not incompetent) by administrators are moved from one school to another in over the years, often coming to rest at the most disadvantaged schools. It also notes the phenomenon of “rubber rooms” – formally called reassignment centers in New York City – where teachers accused of classroom misconduct are charged with minor administrative duties while they await dismissal proceedings. The film provides little context to evaluate the prevalence of these issues, it does point correctly to a sense of outrage in education reform circles over an inability to remove obviously ineffective teachers.

\(^{138}\) As discussed briefly in the introduction (footnote 3, pg. 3), it is ironic that after decades of attempting to standardize teaching methods, reformers turn to grumble about the similarity of teacher effectiveness ratings.

\(^{139}\) The report’s layout does not lend itself to systematic reading; charts and tables are presented in strange formats; some statistics are repeated in multiple places. The endnotes often refer to very substantial qualifications of statements. For example, expanded surveys were administered to four of the school districts (two in Arkansas and one each in Illinois and Ohio); these expanded surveys asked respondents about their specific experiences with evaluation and to report if they knew of tenured teachers who delivered poor instruction (Weisberg, Sexton, Mulhern, & Keeling, 2009, p. 32). These limited subsamples are used to support some of the reports boldest claims.
could improve their teaching. Finally, most surveyed teachers reported that there are poor-performing tenured teachers at their schools. (pp. 13-18)

The report then turns to consider the evaluation systems in the surveyed states. In this portion, they note that most teachers are subject to very little classroom observation time and that there is little relationship between observation time and the ratings teachers received in prior years (such that poorly rated teachers were observed no more than highly-rated teachers). Other statements made in this section are far too sweeping given the data behind them. Most glaringly, the authors conclude this section by suggesting that evaluation policies create a dysfunctional school culture among teachers where feedback is not valued (p. 23).

Widget finally turns to makes policy recommendations for reversing the widget effect. The first is to “adopt a comprehensive evaluation system that fairly, accurately, and credibly differentiates teachers based on their effectiveness in promoting student achievement.” The next three ask that administrators be trained to and held accountable for using such an evaluation system and that these policies be linked to teacher assignment, retention and efficient dismissal (pp. 27-30). While the authors do note in a lengthy sidebar that Value-Added methods could be viable as state testing systems evolve, they stop short of recommending VA outright. They instead stress stronger observation and feedback systems for teachers, rewards for effectiveness and sanctions for poor performance.140

140 What is missing from Widget is a consideration of the drastic changes that would need to be made to teachers’ and administrators’ working conditions to facilitate their recommendations. That administrators spend so little time observing or conferring with their teachers, and that teachers spend so little time collaborating on best practices, are themselves products of an underpaid and overburdened school labor force.
Widget’s basic finding is both valid and startling; it warranted and received much attention. Although much of it relied on somewhat dubious surveys and reporting methods, it addressed a fundamental question. While the report did not recommend value-added methods unequivocally, the alarm it raised on the state of existing systems made a strong case for upending the whole enterprise of teacher evaluation. It is in this context that we ought to understand the mandate in Race to the Top to tie teacher evaluations to student achievement.

These post-NCLB studies bring us to the point where we started in the beginning of this project. What is clear is that value-added research expanded greatly in scope as NCLB testing data became available. At the same time, greatly enhanced computing power facilitated the digitization and statistical processing of larger and more sophisticated educational data. Furthermore, the focus of school reformers shifted sharply toward teacher effectiveness. But value-added research in this period also set the stage for two even larger value-added studies – Long Term Impacts of Teachers and Measures of Effective Teaching. But given their centrality to a widely publicized court case focused on teacher tenure, these will be examined in the final chapter.
Chapter 6: Value-Added in Public: Evidence and Ideology in *Vergara v. California*

Value-Added Assessment was originally conceived as a tool for academic research. However, since the turn of the 21st Century and particularly since the rollout of the 2009 *Race to the Top* program, this method of assessing school and teacher quality has become part of the national debate over education reform, and has been proposed as an evaluation policy. In the previous chapters, we’ve considered how the assumptions of VAA conflict with the practical realities of schooling, how innovations in evaluation have been applied to education historically, and the development of VAA from a research tool to a policy option. To explore the role value-added plays in educational politics beyond its role in research, I examine a recent court case – *Vergara v. California*. The case featured testimony from expert witnesses whose research relies on value-added assessment. The goal here is to gain perspective on how VAA is interpreted outside of research circles.

I begin by outlining some of the details of the case as well as key figures inside and outside the courtroom, focusing on the larger political forces they represent. Then I compare the academic work of key expert witnesses alongside their testimonies in the case. Finally, I assess press coverage following the decision.

Background on the Trial

In 2014, perhaps the most notable political platform for Value-Added Assessment was the proceedings and the decision handed down in *Vergara et al. vs. California*. The case had nine students as plaintiffs against the state department of education and three school districts in California (Los Angeles Unified School District, Oakland Unified School District, and Alum
Rock Union School District). The plaintiffs charged that the state’s teacher tenure policies - particularly the two-year evaluation period before tenure is granted, due process protections related to dismissal, and “Last-In, First-Out” policies - left the districts’ poorest students with the worst teachers, thereby constituting a violation of the equal protection clause of the US Constitution. The California Federation of Teachers (CFT) and California Teacher’s Union (CTU) intervened as defendants, citing their interest in a case involving teacher tenure. Before the trial began, the three school districts were dropped as defendants. The suit was filed in 2012; the trial took place over two months from January 27th to March 27th, 2014. On June 10th, Judge Rolf Treu ruled in favor of the plaintiffs, citing evidence that ineffective teachers harmed the educational and economic well-being of their students. The judge’s 16-page opinion included references to the 1954 Brown vs. Board of Education decision, arguing that equal access to high-quality teachers is among the fundamental educational civil rights issues of our time. The ruling was immediately stayed, preventing it from having any impact on teacher tenure policy until appeals by the defendants are complete. Those appeals were heard in April of 2016.

Although the named plaintiffs in *Vergara* were nine California public school students, the case was the product of an organization called Students Matter, which paid for the plaintiffs’ legal fees and the public relations team that handled the case. The organization’s website describes it as “a national non-profit organization dedicated to sponsoring impact litigation to promote access

---

141 It is important to note that, according to the procedural history in the judge’s decision, the districts were dismissed “voluntarily with prejudice,” meaning that the plaintiffs (Students Matter) dropped the charges against the districts. The superintendents of these districts became plaintiffs’ witnesses.

142 http://cacs.org/research/vergara-v-california-case-context/

143 The unions’ appeals led to the overturning of the Vergara decision. The plaintiffs appealed that decision, but the California Supreme Court refused to hear any further appeals.
to quality public education. The group was founded and is funded by David Welch, an entrepreneur from Silicon Valley.

Students Matter hired the corporate law firm Gibson, Dunn & Crutcher to argue the case; the three co-lead attorneys were Theodore Olson, Theodore Boutrous, and Marcellus McRae. The firm, which ranked fifth nationally in a survey of corporate legal directors, is representing Chevron in an environmental harm case against the nation of Ecuador, and has represented the automobile industry against the Environmental Protection agency in both California and Massachusetts. It also represents New Jersey Governor Chris Christie in the scandal stemming from lane closures on the George Washington Bridge. As a member of the firm, Olson represented George W. Bush in 2001 dispute over the Florida election tallies and subsequently served as US Solicitor General; Olson represented Citizens United in the Citizens United vs. Federal Election Commission case which declared that political spending constitutes free speech. Boutrous was the lead counsel for Walmart in their 2007 gender discrimination lawsuit, which was argued before both the Ninth Circuit Court and the US Supreme Court. McRae is noted for defending corporations against workplace discrimination and environmental pollution lawsuits.

---

144 http://studentsmatter.org/
145 Olson was named as the plaintiffs’ lead attorney, but did not participate actively in the trial. McRae handled the bulk of the proceedings; Boutros offered opening and closing statements. There were a number of other plaintiff attorneys who handled questioning of individual witnesses. The legal team for the defense was much smaller, consisting of one attorney each for the California public schools and the California Teachers’ Union.
146 http://studentsmatter.org/legal-team/theodore-b-olson/
147 http://www.washingtonpost.com/wp-dyn/content/article/2011/01/20/AR2011012005149.html
148 It is notable that the Walmart case also involved expert witness testimony from a statistician and a sociologist.
149 http://www.gibsondunn.com/lawyers/mmcrace
*Vergara* is not the first lawsuit to dispute teach tenure protections in California. In 2010, *Reed v. California* challenged the Last-In- First-Out policy. That case was filed by the Southern California American Civil Liberties Union (ACLU). However, in that case, the plaintiffs were able to establish specific harm to specific students in three middle schools where teacher turnover was particularly high. *Reed* ended in a settlement which was later nullified by the state appeals court for not allowing sufficient input from the teachers’ union.150 *Vergara* is also not the first case to make an issue out of test-based teacher evaluations in California. In the 2012 case *Doe v. Deasy*, Judge James Chalfant ruled that the LAUSD had to make students’ test performance, and particularly Value-Added Models, part of teacher evaluations in compliance with the state’s 1971 Stull Act.151 Similar to *Vergara*, *Doe* was filed on behalf of a group of parents by EdVoice, a nonprofit founded by Netflix CEO Reed Hastings152 which strongly supports charter school expansion and received major support from the Broad Foundation.

The ruling in *Vergara*, whether or not it withstands the appeal by both the National Education Association (NEA) and the California Teachers’ Union (CTU), may set off a series of similar cases in other states. Students Matter President David Welch has promised to promote similar cases in other states, and former CNN personality Campbell Brown has already begun filing a similar case in New York. Brown’s organization, the Partnership for Educational Justice,

---

151 http://www.scribd.com/doc/96728121/Doe-v-Deasy-Tentative-Ruling-Compact-PDF The ruling left it up the school district to negotiate with the teachers’ union over how test scores were to be factored in to teacher evaluations, but as of 2015 no deal was reached (Fensterwald, 2015).
152 Hastings is also a founding member of Green Dot Charter schools, one of the largest charter school management companies in the state.
includes executives from Students First and Democrats for Education Reform on its advisory board.\textsuperscript{153}

This list of expert witnesses’ affiliations in \textit{Vergara} further indicates the interlocking network of private and non-profit interests that has gained power in US education policy. The key experts for the plaintiffs – those whose testimonies featured in the judge’s ruling – were three economists: Harvard’s Thomas Kane and Raj Chetty and the Stanford’s Eric Hanushek. Kane was the principal investigator for the Gates’ foundation-funded “Measures of Effective Teaching” (MET) study and is a strong advocate of using Value-Added Assessment to evaluate individual teachers’ effectiveness in the classroom. Raj Chetty was called as a witness based on his study of the long-term impacts of effective teachers. Eric Hanushek, also a senior fellow at the conservative Hoover Institute, has been writing about performance-based teacher and school evaluation since the 1970s, and has further argued that input-based schooling policies – those which focus on resources and teacher credentials – have little impact on educational quality. Hanushek was also a member of former California Governor Arnold Schwarzenegger’s Committee on Education Excellence, which received support from the Gates Foundation and whose other members include executives of Students Matter and the New School Venture fund – another corporately-funded non-profit organization.

Other plaintiff experts included Dan Goldhaber, a labor economist at the University of Washington – Bothel; Arun Ramanathan of EdTrust – West\textsuperscript{154}, a non-profit organization which lists “Accountability” and “Teacher Effectiveness” as two top items on their advocacy agenda.

\begin{footnotesize}
\begin{itemize}
\item \textsuperscript{153} http://www.edjustice.org/about/board-directors-advisory-board/
\item \textsuperscript{154} As of April 2014, Ramanathan is now the CEO of Pivot Learning Partners, a company that works on the implementation of Common Core Standards.
\end{itemize}
\end{footnotesize}
and which has received funding from both the Gates and Walton foundations; and Sandi Jacobs, vice-president of the National Council on Teacher Quality (NCTQ) and a former program specialist with the US Department of Education. NCTQ receives funding from the Broad, Gates and Walton foundations; its advisory board includes former NYC schools chancellor and current News Corporation VP Joel Klein, founder of KIPP charter schools Michael Feinberg, Pearson executive Michael Barber, and fellow expert witness Eric Hanushek (National Council on Teacher Quality, nd).

Among its experts, the defense called one economist, Jesse Rothstein of the University of California-Berkeley. Rothstein has been a vocal critic of value-added assessment, arguing among other things that its statistical assumptions are never met in practice in actual schools. He also wrote a critical analysis of the MET study findings for the National Education Policy Center (NEPC), where he is a fellow. The other expert witnesses for the defense were all education researchers: Linda Darling-Hammond of the Stanford Graduate School of Education; Susan Moore-Johnson a colleague of Kane’s at Harvard; Ken Futernick of California State University – Sacramento; and David Berliner, Professor Emeritus at Arizona State University and a researcher at NEPC.

Other key witnesses were Oakland Schools Superintendent Tony Smith and Los Angeles Schools Superintendent John Deasy. Both Smith and Deasy testified for the plaintiffs, despite initially being named as defendants in the case. Their testimony is not grounded in research, but becomes important when we consider the press coverage of Vergara.
Contextualizing Foundation Influences

Students Matter is a relatively small player among nonprofits interested in education reform. But David Welch’s effort in *Vergara* was supported directly and indirectly by many of the largest players in educational “venture philanthropy." The Broad Foundation was one of the earliest contributors to Students Matter. The David and Heidi Welch Foundation has made significant donations to NewSchools Venture Fund, which invests in both traditional and online charter schools, and to StudentsFirst, the education reform group founded by former Washington, DC schools’ chancellor and tenure opponent Michelle Rhee.

Foundation influences are also evident among the school districts which were dropped as defendants. John Deasy, the Los Angeles School Superintendent and plaintiff witness, served as a director for the Gates Foundation and was a Broad Foundation Fellow. Former Oakland Superintendent Tony Smith, a Students Matter advisory board member who also testified for the plaintiffs, was the director of the National Equity Project which, under its former name (Bay Area Coalition of Equitable Schools), was a recipient of Gates Foundation support for its emphasis on “small schools.” Smith is now executive director of the Stone Foundation, which has made education grants to groups such as the Academy for Urban School Leadership – a school management organization which orchestrates school turnarounds and receives support

---

155 This term has been used by Diane Ravitch to describe a sort of 21st century philanthropy related to education. The term is described in her 2010 book *The Death and Life of the Great American School System,* where she refers to a particular group of foundations including the Bill and Melinda Gates Foundation, the Eli and Edythe Broad Foundation, and the Walton Family Foundation as the “Billionaire Boys Club.”
156 http://home.lausd.net/apps/pages/index.jsp?uREC_ID=178743&type=d&pREC_ID=367410
158 http://classroomstruggle.org/2013/04/08/tony-smith-leaves-oakland-in-shambles/
from both the NewSchools Venture Fund and the Gates Foundation. He also heads Teach Plus, an advocacy organization for “reform-minded teachers” which promotes test-based teacher evaluations and merit-based compensation plans and receives support from both the Gates and Broad foundations.

Diane Ravitch (2010) notes the growing influence of what she calls the “Billionaire Boys club” of large philanthropic foundations. Ravitch points out that after a long hiatus following the ‘School Wars’ of the 1960s, large philanthropies returned to education reform in the late 1990s, but with a radically different program. In contrast to the older foundation approach – which allowed grantees to spend grant money on their own priorities – new ‘venture philanthropists’ set their own agendas and sought out organizations and individuals aligned with their goals. She further points out that as the three largest education-focused foundations – the Bill and Melinda Gates Foundation, the Edith and Eli Broad Foundation, and the Walton Family Foundation – converged on similar priorities (along with many smaller funders), cash-starved school districts as well as state and federal education agencies began to align their agendas with those of the foundations. Ravitch problematizes this state of affairs:

“There is something fundamentally antidemocratic about relinquishing control of the public education policy agenda to private foundations run by society’s wealthiest people; when the wealthiest of these foundations are joined in common purpose, they represent an unusually powerful force that is beyond the reach of democratic institutions…The foundations demand that public schools and teachers be held accountable for performance, but they themselves are accountable to no one.” (Ravitch, 2010, pp. 200-201)

Although focused on India, Arundhati Roy (2012) gives an interesting historical perspective on corporate philanthropy in the US. She notes that when the first corporately-endowed foundations

---

159 http://auslchicago.org/about
160 http://www.teachplus.org/page/partners-27.html
– the Carnegie Corporation and the Rockefeller Foundation – appeared in the early 20th century, critics raised serious questions about the legality and unaccountability of such institutions. Roy connects early corporate philanthropy with attempts to sway US foreign policy and to influence domestic policy in developing countries in ways that would favor corporate profits. Like Ravitch, she also notes the irony of opaque, unaccountable organizations demanding transparency from foreign governments.

Picciano and Spring (2013) offer detailed description of this phenomenon in American education in their description of the ‘Education-Industrial Complex.’ They argue that control over education policy has become dominated by a network of individuals who move between posts in government, non-profit foundations and for-profit firms. These ‘flexians’ use their influence to promote school reforms which emphasize private sector partnerships and the application of technology to education. As such, they tend to oppose teachers’ unions and support top-down, data-driven policies like Value-Added assessment. Picciano and Spring argue that the growing influence of foundations and the increasing privatization shift power away the democratic processes that have historically governed public schools. Instead, democratic governance is replaced by a ‘personalized bureaucracy’ in which school policy is determined by a closed group of individuals on a first-name basis with each other (Picciano & Spring, 2013, p. 38).

As the above description of Students Matter and the roster of expert witnesses indicate, the tendencies toward privatization and foundation influences were on display in the Vergara v. California trial. We now turn to consider the testimonies of expert witnesses in the trial alongside the research on which they were based.

---

161 Picciano and Spring borrow the term flexian from Janine Wedel (2009).
Research Basis of Two Expert Witnesses\(^{162}\)

Raj Chetty was the first expert witness, called by the plaintiffs on the third day of the trial.\(^{163}\) Chetty’s co-authored paper with John Friedman and Jonah Rockoff focuses on the measurement of teacher quality and the long-term impacts of effective (and ineffective) teachers.

The research for these papers first took shape as an Internal Revenue Service white paper on the impacts of income tax credits. Using New York City school records from the 1990s linked with tax records, that paper found that a $1,000 increase in the amount of a family’s tax credit was associated with an increase in student test scores between 6 and 9 percent of one standard deviation.\(^{164}\) They then correlated test scores with earnings in early adulthood. These results showed that a one standard deviation increase in student test scores in a single grade correlated with 9 percent increase in earnings at age 28. Combining these findings, they suggest that tax credits are more than offset by the long-term earnings gains they afford (Chetty, Friedman, & Rockoff, 2011).

The IRS paper evolved into a National Bureau of Economics working paper which shifted the focus to teachers. The first part of the *Long Term Impacts* paper is devoted entirely to the question of bias in Value-Added measures – whether VA estimates are driven by the causal impact of teachers or by student sorting. First, the authors devise a novel test using the concept

\(^{162}\) Eric Hanushek’s role as the architect of Value-Added Assessment has already been considered in the preceding chapters; I also consider his testimony below.

\(^{163}\) Video of Chetty’s testimony is available online at http://vimeo.com/85786192

\(^{164}\) All test score gains cited in the various iterations of Chetty et al. are phrased in terms of standard deviations rather than as raw test scores. This is because in the two decades of data they use, the school district (New York City) underwent numerous testing changes which make the raw test scores incomparable. Indeed, as I note, in an earlier chapter, changes in testing instruments are one outside factor which compromises the validity of test-based teacher evaluations.
of forecast bias.\textsuperscript{165} They correlate the changes in grade average teacher value-added with the changes in grade average test scores. They find a very high correlation between these two variables.\textsuperscript{166} They find that estimates of teacher value-added in a given year are not affected by the addition of other highly predictive variables (parent characteristics and lagged test scores). Their second (and more robust) test – a quasi-experiment analyzing teacher turnover events – found that when a high- or low-value-added teacher arrives at or departs from a school, the average effect of same grade-level teachers on student test scores can be considerable and vary in the direction of her estimated value-added – e.g., a high-VA teacher positively impacts her grade’s average test scores when she arrives. These findings lead them to conclude that teacher value-added is an accurate predictor of a teacher’s contribution to student test scores.

\textsuperscript{165} Forecast bias is a very technically elaborate procedure in the case of measuring teacher effects. The method begins by conceiving of changes in teaching staff as experimental treatments administered to a grade full of students. The hypothesized effect is derived from a jackknife estimate (leaving out the year before and the year of the change) of teachers’ combined impacts on student test scores in all other available years. So if a grade of teachers had student test score data for 10 years – say 1991 through 2000 – and one wanted to hypothesize an effect for changes between 1993 and 1994, one would average the data from 1991, 1992, and 1995-2000, leaving out the years in question, which the authors argue are measured with error that is related to student outcomes (Chetty, Friedman, & Rockoff, 2011a, p. 21). This seems suspect because it uses data from subsequent years in a teacher’s career to estimate her effect on students in the past. In the real world, schools would not be able to use a teacher’s future performance to predict her current ability. All the averaging of effects across grades is also not aligned with how value-added is being used in policy.

\textsuperscript{166} The figures presenting their estimates of forecast bias make use of “binned scatterplots” – for which the authors helped developed new software. This technique takes the thousands of points (in this case 60,000) that would be plotted on a graph and groups them into a user-determined number of bins (in this case 20), which are then assigned the average values of the x and y variables depicted on the axis of the graph. These binned averages become points on the scatterplot. Binning reduces the cluttered appearance of scatterplots based on large datasets. It can also help easily communicate the functional form of the relationship between variables. But binned scatterplots do not communicate the strength of the relationship between two variables because they tell us nothing about the R-squared value or proportion of variance explained (Stepner, 2014). Further, the visual effect of reducing a graph that would have thousands of points to one with only twenty data points is important; it makes relationships appear much clearer.
The second part of the paper utilizes the unique features of the data to assess the long-term impacts of teacher quality – as measured by value-added – on students. They begin with the finding that a one standard deviation increase in teacher value-added increases students’ test scores by 10 percent of one standard deviation. This is an important step because the long-term impacts findings are based on the correlation between those outcomes and a one standard deviation increase in student test scores, not in teacher VA – a one SD improvement in student test scores would necessitate an impossible ten standard deviation improvement of teacher VA. So the authors simply divide their findings by ten to approximate the impact of a one SD improvement in teacher value-added. The primary outcomes discussed are college attendance at age 20 and earnings at age 28. The authors report that on average, a one standard deviation increase in teacher value-added in a single grade increases a student’s probability of attending college by 0.49 percentage points, and increases student earnings at age 28 by $182. These are modest findings given that, in terms of policy, a one standard deviation increase in teacher value-added would require schools to affect a 34 percent increase in overall teacher quality, however defined.

These findings are accompanied by various extrapolations which require strong assumptions about how teachers are related to test scores, and how test scores are related to long-term outcomes. For example, the $182 earnings gain is extrapolated to a policy of “deselecting” (read firing) the bottom five percent of the value-added distribution from the teaching force and replacing them with average teachers.\textsuperscript{167} They argue that such a policy would increase each

\footnotesize
\textsuperscript{167} The authors note that teachers in the bottom five percent of the distribution are on average only two standard deviations below the mean, suggesting that the distribution of teacher value-added is clustered closer to the mean. Their use of the bottom five percent follows from Hanushek’s earlier research on teacher replacement. They also assume in their calculations that the replacements would not be new teachers, but those with average experience (Chetty,
student’s lifetime earnings by $9,422 – but under difficult assumptions including that: a) higher-VA teachers affect all students in the same way, b) earnings gains would be plausible if all students got higher-VA teachers, and c) that replacements of average quality are readily available at any time.

*Long Term Impacts* received much attention even as a working paper – including front-page coverage in the *New York Times* and significant mention in President Obama’s 2012 State of the Union Address. Still, many authors have raised questions about the findings of Chetty et al. in this study. Richard Rothstein (2012) suggests that the focus on teachers’ abilities to raise test scores elides other important aspects of teaching. Combining the findings of *Long Term Impacts* with those of the *Measures of Effective Teaching* study (discussed below) he contends that this narrow focus ignores other ways teachers impact students which could in fact have more dramatic long-term impacts.

Dale Ballo (Ballou, Review of The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood, 2012) argues that although the authors find no forecast bias in the estimation of teacher value-added, they fail to address this same type of bias in their assessment of long-term impacts:

> “Even if we could be assured that the value-added model perfectly measures a teacher’s contribution to student knowledge [test scores], it would remain problematic to attribute long-term outcomes to the impact of teachers… [it must be established that] high value-added teachers are not more likely to have been assigned students who were for other reasons destined for greater long-term success.” (pp. 5-6)

Just as the authors tested for bias in value-added measures by checking for the impacts of parent characteristics, the same tests should be used to establish that the effects of teachers on long-term

---

Friedman, & Rockoff, 2011a, p. 48). This assumption can be made in theory, but is unlikely to hold in the actual teacher labor market.
outcomes are not being confounded with the impacts of parental characteristics on earnings, college attendance, etc. Ballou suggests that without these tests for bias in the prediction of long term outcomes, not just test scores, the validity of the findings of *Long-Term Impacts* is called into question.

Taking a different approach, Moshe Adler (2013) looks closely at what is actually measured in Chetty et al.¹⁶⁸ He raises two major issues with the findings. The first concerns the lynch-pin linking the two studies. As noted above, the thread that connects the authors’ conclusion of unbiased measurement of value-added (part one) and the long-term impacts of teacher-value added (part two) is that a one-standard deviation increase in student test scores is shown to correlate with long-term outcomes. There are two problems with this connection. The first is the fade-out of teacher effects on test scores, which Adler documents from numerous prior studies (2013, p. 5). If test score effects don’t last, then the correlation between test scores and long-term outcomes is improperly attributed to the effect of teacher quality. The second more serious problem is the measurement of value-added at the teacher level and the student level. A one-standard deviation increase in teacher value-added is shown to increase students’ test scores by 10 percent of a standard deviation. Thus, to increase student test scores by a full standard deviation would require increasing *teacher value-added* by ten standard deviations. Though the authors reduce the magnitude of their reported effect size to reflect this, the problem is the effects they actually find are based on changes in teacher quality that are nearly impossible.

Adler further supports his critique by noting that the range of teacher value-added documented in *Chetty et al.* is between -0.18 and 0.18 standard deviations of student test scores – reflecting that

¹⁶⁸ Adler is unconvinced by the critiques offered by Rothstein and Ballou. He contends that these critiques of Chetty et al. do not decisively undercut the authors’ claims because they focus on what is missing from their argument rather than the strength of the evidence offered.
teacher effects on student test scores are clustered around the mean, adding to the implausibility of assuming that teachers could account for massive changes in student test results.\textsuperscript{169}

The second broad aspect of Adler’s critique is the selective focus on outcomes. \textit{Long-Term Impacts} finds a statistically significant impact of teacher value-added on student earnings at age 28. They also measure earnings at age 30, but state that “the 95 percent confidence interval for the estimate [of earnings at age 30] is very wide.” This is a dubious way of saying that the impact on earnings at age 30 was not found to be statistically significant (Adler, 2013, pp. 3-4).

Unfortunately, one would need to read deep into the paper to discover that; appendix table 6 shows that the point estimate is $206 with a standard deviation of $195.\textsuperscript{170} Nonetheless, the authors assume that the earnings gains found at age 28 would persist when they estimate both the lifetime earnings impact of high value-added teachers and the impact of deselecting the bottom five percent of teachers. In sum, Adler finds that the evidence offered in \textit{Long-Term Effects} is thin given the scope of its claims.

Thomas Kane was the second expert witness called by the plaintiffs in \textit{Vergara}. His testimony focused first on the results of the Measures of Effective Teaching project (funded by the Gates Foundation), for which he was principal investigator\textsuperscript{171}. He also testifies based on his study of the distribution of ‘bad teachers’ in Los Angeles in terms of students’ ethnicity. While the MET study report stresses the need for multiple measures of teacher ability, value-added assessment

\textsuperscript{169} This range is cited in numerous figures in the \textit{Long-Term Impacts} paper (Chetty, Friedman, & Rockoff, 2011a, pp. 60-67).
\textsuperscript{170} Chetty et al. claim that the sample size of 30-years olds is too small to generate a reliable estimate, but Adler argues that over 61,000 student records provide more than enough statistical power (Adler, 2013, p. 4).
\textsuperscript{171} In addition to Kane, the other authors of the MET study report were Daniel McCaffrey and Trey Miller from the RAND Corporation, and Douglas Staiger of Dartmouth College (Kane, McCaffrey, Miller, & Staiger, 2013).
serves as the study’s benchmark quality metric. Further, the latter portion of Kane’s testimony relies solely on Value-Added measures of and specifically his replication of Chetty’s argument about long-term impacts. But given the prominence of the MET study, it is important to consider in its entirety. I focus here on three MET study reports.

The MET project brought the weight of randomized control trials (RCTs) to bear on the question of test-based teacher assessment. Since the passage of the Education Sciences Act in 2002, RCTs have been regarded as the gold standard for educational research. The Gates Foundation has been particularly important in funding education-focused RCTs – the foundation funded the MET project for $45 million. The MET RCT included over 1,000 teachers across more than 300 schools in six geographically diverse school districts. It attempted to fulfill one assumption of value-added assessment which is never met in real school settings: random assignment of students to teachers. Randomization of students had to take place within school, grade and subject – requiring a complex process assignment procedure. In addition to existing state

---

172 In the first of the MET study’s reports, the authors lay out three premises including that “whenever possible a teacher’s evaluation should include his or her students’ achievement gains” and that “any other components of the evaluation should be demonstrably related to student achievement gains” (Kane & Cantrell, 2010, pp. 4-5). Rothstein (2011) notes that these assumptions are problematic for a project studying effective teaching – as they rule out in advance the possibility that test score gains may be a poor measure of teaching ability. Baker (2013) has a similar critique, noting that using Value-Added as the outcome measure of quality makes it almost certain that prior Value-Added will be the best predictor of quality. Goldstein (2015) connects these assumptions to the Gates Foundation’s already-stated policy priorities and to its close ties with the Obama Administration, which had already endorsed the practice of test-based teacher evaluations.

173 The initial pool of teachers was over three thousand, but attrition by both teachers and students whittled that number down to 1,181 (Kane, McCaffrey, Miller, & Staiger, 2013, p. 45). Nonetheless, the MET study is still the largest educational experiment ever conducted (Rothstein & Mathis, 2013, p. 5).

174 Rosters of students were randomized among a block of teachers who taught the same grade and subject in a school. For example, students would be randomly assigned among three fourth grade teachers in a participating school.
tests, MET employed a number of new instruments to measure teacher ability in different ways (project-specific achievement tests, student surveys and standardized assessment of teaching practices).

The first portion of the project focused on correlating value-added assessment with other measures of teaching using non-experimental methods. This was also the first stage of the RCT; teachers who agreed to participate in the study were ranked (according to their value-added scores in 2009-2010) as ‘more effective’ or ‘less effective’ prior to being randomly assigned a roster of students in 2010-2011.

Among their findings, the report highlights that teacher value-added is relatively stable across classrooms and across different testing instruments. Correlations between two sections taught by the same teacher in the same year ranged between .09 and .38 depending on the subject and instrument used. For elementary school teachers who do not teach multiple sections, value-added in one year correlated modestly with the prior year (.40 for Math and .20 for English).

Correlations between the state tests and the MET-administered tests were .377 for ELA and .221 for Math (Kane & Cantrell, 2010, pp. 23-25). The authors adjust these last figures by exploiting data on teachers who have multiple sections to estimate the correlation between ‘persistent teacher value-added’ on the different assessments – which are reported as .54 for math and .38 for English Language Arts (ELA) (Kane & Cantrell, 2010, pp. 18-21).

This first report concludes by using non-experimental methods to compare teachers with the least evidence of effectiveness (measured by value-added) to those with the most. This analysis is

---

175 This report required careful reading. Though the text of the report claimed these correlations were listed in Table 5, they were in fact reported in tables 7 and 8 (Math and English were reported separately).
confined to those teachers who taught more than one classroom in 2009-10, as it uses value-added ranking in one class to predict value-added in the other. They report differences between teachers in the top and bottom quartiles of the effectiveness distribution and find between a 0.07 and 0.33 standard deviations in average test score gains between these two quartiles (the magnitude of difference varies by instrument). For ease of interpretation, the authors convert these differences to estimated months of learning based on a scale developed by Kane (2004). Using a “rule of thumb” They report that the average difference between high-effectiveness and low-effectiveness teachers is between 2 and 12 months of learning (Kane & Cantrell, 2010, p. 27). The findings from this analysis serve as the precursor to the random assignment study.

The MET project’s final report, Have We Measured Effective Teaching?, was released in January 2013, and features the results of the random assignment study. The report begins by describing the algorithm used to rate teachers prior to randomization – which includes 2009-10 value-added, classroom observation ratings, student survey scores, and teacher characteristics (Kane, McCaffrey, Miller, & Staiger, 2013, pp. 10-11). It then demonstrates that the observed growth gains were close to their predictions using a binned scatterplot.

From a methodological standpoint, the most significant outcome of the study may be its high rates of attrition and non-compliance. The original intent was to randomly assign students to

---

176 The authors acknowledge that this scale is based on only one outcome measure, and that this conversion to months is not appropriate for the two state tests, which are not scaled vertically (Kane & Cantrell, 2010, p. 27).

177 Following the general logic of the MET study – where score gains are the priority – the outcome that these factors predict is test score growth. Teachers with the highest predicted average test score growth are those designated highly effective.

178 This is a consequential presentation choice. Teachers were sorted into groups of 40, dramatically reducing the number of points on the regression line and thereby hiding the amount of divergence from the line (Kane, McCaffrey, Miller, & Staiger, 2013, pp. 4-5).
2,642 teachers using a very complicated process and requiring a number of preconditions. 770 teachers were either not eligible for randomization because of job changes; 281 were dropped from randomization because they had did not have another teacher in the same school-grade-subject with whom to exchange rosters; a further 441 teachers were excluded because the school in which they taught was excluded from the study. Thus, only 1,181 teachers were randomly assigned students. But the random assignment was further challenged by subsequent non-compliance. Across all six participating districts, between 34 and 73 percent of randomized students did not remain with their assigned teacher. Further, non-compliance appeared to be systematic: teachers who were part of randomization had students with higher average test scores, were more likely to be white, less likely to be African American, and less likely to be classified as either special education students or English-language learners (Garrett & Steinberg, 2015). The authors examine and acknowledge that even among randomized teachers, there was some sorting of students to teachers – suggesting that student roster changes made after randomization were not themselves random (Kane, McCaffrey, Miller, & Staiger, 2013, pp. 16-17). They also note that because students were not randomly assigned to schools, the findings are limited to whether value-added measures are biased within schools, and cannot say whether value-added was biased between schools. Using a final sample of 27,255 students, they proceed to assess the question: Does a one unit increase in predicted teacher effectiveness translate into a one-unit difference in observed student achievement?

179 Steinberg and Garrett (2015) re-analyzed the MET data and found that indeed this sorting has a significant impact on measured teacher performance, following Rothstein’s (2010) criticism of value-added models.

180 As my analysis in Chapter 2 shows, school level variables are important determinants of teacher value-added scores, suggesting that between-school bias may be significant.
Thus the MET project uses random assignment to assess the treatment effect of teachers. However, owing to the high non-compliance rate, the analysis relies on an alternative method of identifying teacher (treatment) effects: Instrumental variables (IV). This shifts their focus to the effect of treatment on those students who remained in the randomization blocks – what the literature on experimentation calls “compliers.” Compliance varied by district; in one municipality, a student’s assigned teacher had no correlation with their actual teacher’s effectiveness. Nonetheless, in their analyses, the authors find support for their hypothesis among the specified population – concluding that the average effect of being assigned to a more effective teacher reflects measures of that teacher’s effectiveness (Kane, McCaffrey, Miller, & Staiger, 2013, p. 38).

After the random assignment study was complete, the MET project also released a paper which addressed the use of multiple indicators in teacher evaluations (Mihaly, McCaffrey, Staiger, & Lockwood, 2013). In addition to scores on state tests and project-specific tests, the MET project

---

181 In RCTs with considerable non-compliance, simple outcome comparisons of the treatment and control groups are regarded as unreliable. IV methods are one approach to identifying treatment effects in such circumstances. IV attempts to estimate the average effect of the treatment on those who actually received treatment. The modern approach to IV was developed by economists working with non-experimental data in so-called natural experiments (Angrist, Imbens, & Rubin, 1996). Researchers working on random assignment studies have since adopted this method (see for example Hedges, 2008). Using IV methods requires that strong assumptions be met especially when dealing with “weak instruments,” such as RCTs with very high non-compliance, like the MET study (Staiger & Stock, 1997; Imbens, 2014).

182 Imbens and Angrist (1994) outline conditions required to adequately estimate local average treatment effects. Among them is the “no-defiers” or “monotonicity” assumption, which states that the treatment needs to affect the outcome in only one direction, such that treatment is always better (or always worse) than control. In the case of MET, where teachers were randomly assigned to classrooms, it seems likely that a high-value-added teacher would not always affect all students in the same direction (e.g., positively impact achievement for everyone). This violates the no-defiers assumption and problematizes their estimation of LATE.

183 They are careful to note that the findings among English teachers are considerably imprecise, but do not mention that results among middle school students were also more imprecise.
conducted student surveys in all participating classrooms and collected multiple videotaped lessons for all participating teachers. Each lesson was rated using general and subject-specific rubrics. These two sources of data were each used to construct standardized scales which were incorporated into composite estimators of teacher quality.

This report, though written in a technical style, provided perhaps the some of the project’s clearest findings. First, the authors report that all three measures of teacher quality (value-added, observation ratings, and student surveys) are somewhat unreliable—the same measure fluctuates from year to year. Second, they show that different instruments of the same ‘mode’ (e.g., observation scores for two different rubrics or value-added measures for two different tests) are far more correlated than with each other than they are with any other mode (e.g., value-added with observations). They interpret this to mean that each indicator is capturing a relatively unique aspect of teacher effectiveness. Finally, they demonstrate that the optimal weighting of a composite estimate of teacher quality depends almost entirely on the outcome one wants to predict. If, for example, one wishes to predict teacher observation scores, the optimal predictor will put preponderant weight on another observation score. Their evidence leads them to conclude that decisions about how to rate teacher effectiveness are not matters of picking the most statistically reliable indicators, but rather of deciding which aspects of teaching are the most important outcomes, and that if one outcome is chosen, little is gained by using a composite

---

184 Goldstein (2015) notes that MET’s protocol of filming teachers at work generated controversy in the media.

185 Looking at the data supporting this point (Table 3 on page 24 of the report), another interpretation is possible. The average correlations between the two subjective measures – student surveys and teacher observations – are somewhat higher than those between value-added and either subjective measure. One interpretation of this is that while observations and surveys address what happens in classrooms between teachers and students, test score gains are determined by a combination of endogenous and exogenous forces.
estimate (Mihaly, McCaffrey, Staiger, & Lockwood, 2013, p. 39). If any composite measure is to be used, the authors suggest equally weighting each of the three modes of assessment. Nonetheless, much of the report focuses on how to optimally weight and reliably predict teacher quality when value-added is the outcome of interest. This focus was chosen based on “the current policy focus on improving student test scores” (p. 25).

While recognizing the scope and importance of the MET study, critics have raised questioned the presentation and interpretation of its findings. When considering the RCT specifically, such high levels of attrition raise serious concerns about generalizability of those results. After problematizing the assumptions laid out in the first MET report [see footnote 28 above], Rothstein (2011) critiques the report’s assertion that different test instruments consistently identify effective teachers. Individual teachers’ between-test correlations averaged .54 in mathematics and .37 in ELA, indicating that many teachers identified as effective by one test would be ranked below average by the other test. Rothstein also challenges the authors’ estimation and correlation of a stable component of value-added, which can be inferred by statistical procedures, but not definitively ascertained. Further, reporting correlations between the stable components of value-added measures overstates the precision of VA estimates (Rothstein J., 2011, pp. 3-5). The MET study’s admittedly imprecise method of converting test score gains into months of student learning has been criticized by as a way of inflating the value of objectively small differences in actual test performance (Rubinstein, 2014; 2013; DiCarlo M.,

---

186 Language is conveniently flexible among MET reports. The authors call a correlation of 0.54 between value-added measures “moderately large” (Kane & Cantrell, 2010, p. 9). However, when correlating value-added with subjective measures – observations and student surveys – 0.53 is called “weakly correlated” (Mihaly, McCaffrey, Staiger, & Lockwood, 2013, p. 28).
2012). A larger criticism of the early MET findings focuses on the difference between high-stakes evaluation and research. The value-added data collected were not used for decisions about dismissal or tenure; if they were, the data would likely be very different. Thus although the study is intended to validate value-added assessment for personnel decisions, it is unable to address that issue (Rothstein J., 2011, p. 7).

The random assignment component of the MET study has also been subject to scrutiny. Rothstein and Mathis (2013) argue that although the study is large in scale, the voluntary opt-in by districts and schools meant that even the intended sample was not representative of the US K-12 system. Participating schools within the six MET districts were on average higher-performing than non-participating schools. Further, student non-compliance with the study was far from random; students who remained in the study as assigned were on average higher-performing than those who shifted from their assigned teacher. The authors of this review point out that the subgroups that were under-studied in the MET experiment are precisely the subgroups which raise questions about bias in value-added assessment (pp. 5-6). Owing to these limitations, they argue that the study does little to improve on Kane and Staiger’s 2008 random assignment study in Los Angeles and does not resolve the question of whether value-added measures are biased based on student sorting. Furthermore, the statistical VA model used in the MET reports is far

---

187 This method of converting average test scores to months of learning can be traced back to a study by Eric Hanushek (1992). It has also featured prominently in research on the effectiveness of charter schools (Center for Research on Educational Outcomes, 2012) and of alternative teacher certification methods (National Center for Educational Evaluation, 2013).

188 The MET project final report analyzes all students who were randomly assigned, even though a substantial proportion (14 percent) of those did not remain with their assigned teacher (Kane, McCaffrey, Miller, & Staiger, 2013, p. 14).
more comprehensive than those being used for consequential evaluations – which further limits the utility of the results (pp. 7-9).

Critics regard the *Composite Estimator* findings as much more interesting, perhaps because they are less controversial. That no two modes of evaluation were highly correlated suggests that decisions about evaluation are necessarily value judgments, not matters of technical precision.

Notably, the What Works Clearinghouse, a federally funded program which assesses the strength of education research and focuses on experimental and quasi-experimental methods, did not assign a rating to the MET study. WWC said that though the study utilized random assignment, it did not “test a specific policy, intervention or practice” (What Works Clearinghouse, 2013). But this is not exactly the case. The MET project poses and provides an answer to an ‘RCT-ready’ question – what is the causal effect of a one unit change in measured teacher effectiveness? So it is interesting that the study was not rated. But the WWC standards suggest an explanation. The review process focuses on Intent-to-Treat (ITT) analyses of effects – which were not provided in the MET reports. And the MET study population differed substantially from the general population on certain key traits (baseline test scores, ELL and Special Education status). For the WWC, this compromises the use of the alternative instrumental variables (IV) approach. Further, in its use an alternative effect measure, the study does not include a baseline equivalency measure of SES, which is a requirement for meeting WWC standards.

---

189 Despite this claim by WWC, it reviewed parts of the Chetty et al. long-term impacts study (What Works Clearinghouse, 2012).

190 Two of the study’s lead authors have used ITT estimates of treatment effects in a smaller study of teacher effectiveness conducted in Los Angeles (Kane & Staiger, 2008). The MET study was an elaboration of the design used in that paper (Rothstein & Mathis, 2013).
standards when there is high attrition. Thus it is likely that if the WWC were to review the MET project as a random assignment study, it would not meet their evidence standards.

Two key witnesses in the Vergara trial – Thomas Kane and Raj Chetty – each made significant, high-profile contributions to value-added research. These contributions feature prominently in their expert testimony. But they are not without significant limitations or criticisms. Thus we now turn to examine the expert testimony to see how their research is translated into trial evidence.

Expert Testimony in Vergara—Raj Chetty

In considering the expert testimony of Thomas Kane and Raj Chetty, I examine transcripts of their testimony and the demonstrative exhibits used in their presentations. The transcripts are uncertified drafts and were accessed from a website established by the California Teachers Union during the trial. The charts and diagrams used by expert witnesses were accessed through the Students Matter website.

Raj Chetty testified in the Vergara trial over the course of two days – January 29th and 30th 2014. He was examined by Theodor Boutros for the plaintiffs and cross-examined by James Finberg and Nimrod Elliot for the defense. The evidence related to Chetty’s testimony was his National Bureau of Economic Research (NBER) Working Paper released in 2011 (Chetty, Friedman, & Rockoff, 2011a) and slides based on that research. The early part of the testimony – 35 minutes – was spent elaborating on Chetty’s career as an economist – including his degrees, other criticisms were made of Chetty et al. and the MET project, but I focused on those which were available before the trial. Chetty’s papers were admitted as plaintiffs’ exhibits 7, 8, 9, 11 and 15. The slides were admitted as plaintiffs’ exhibit 677 (p. 582).
awards, involvement with government officials, and research interest in public economics (p. 436-455). Before proceeding into his testimony on value-added measurement, Chetty also connected VA research to the use of “Big Data” in solving policy problems. He justifies this as promoting a scientific and non-ideological approach to social policy – perhaps not considering the fact that his role as an expert witness in a trial seeking to undermine teachers unions is a profoundly ideological choice (p. 461).193

His testimony on value-added methodology focused on so-called intuitive concepts rather than technical details. For example, measuring gains in learning by subtracting average fourth grade test scores from average third grade test scores (p.478) is an intuitive method. But [as I showed in chapter one] it is a technically fraught approach to isolating an individual teacher’s effectiveness – since there are effectively two teachers’ effects being estimated under a typical testing schedule. Chetty’s discussion of demographic controls or ‘fairness variables’ in VAA argues that prior year test scores are the only important control (p.482), a claim which is still in dispute among VA researchers, especially given that non-random student sorting is prevalent in most schools districts (cites).

The discussion then shifts to the quasi-experimental tests for bias reported in the Long Term Impacts paper. In explaining how teacher turnover creates a natural experiment, he refers to random assignment as “a gold-standard scientific experiment” (p.488). This presages the testimony given later by Kane, but also belies the trouble with compliance in such studies. He

\[193\] “So at a high level what I’m interested in as an economist, what motivates me is trying to bring a scientific approach to bear on public policy issues. I think too often in our public policy debate; decisions are made on the basis of ideology or political interests rather than what I think can be really scientific answers.” This is consistent with what Chetty argues in a New York Times Op-Ed (Chetty, 2013).
then describes the quasi-experimental observations of moments in which a highly effective teacher (top 5 percent of the VA distribution) enters a grade. Chetty’s description of his own findings at this point is inaccurate – he attributes the growth in test scores solely to the highly teacher effective teacher (pg. 489). On the other hand, the Long-term impacts paper itself reports that this is the growth in average test scores across the entire grade, not just in the classroom of the newly hired instructor (Chetty, Friedman, & Rockoff, 2011a, p. 61). This is a consequential difference because it is not necessarily clear that the newly hired teacher is on her own responsible for the growth – one could imagine that instructors can and do collaborate, especially when a staff member is new to a school. The same point can also be made in reference to the next few minutes of his testimony, which concern the entry and exit of highly effective or ineffective teachers (p. 490-500) The year Chetty analyzes is also of consequence – New York City (the data source) changed its reading test instrument between the 1995 and 1996 school years. As a result, the proportion of students who scored at or above grade level in reading dropped from 47.5 percent to 41.6 percent – leading local officials to caution against over-interpreting the change in test scores (Newman, 1996).

194 “Now what happens as soon as that teacher enters, test score immediately jump up, and they continue to stay high for subsequent cohorts of children who are taught by this highly effective teacher [emphasis added]” Chetty is clear that these figures are taken from the Long Term Impacts paper, which is listed as exhibit 7 for his testimony (p. 492). He clarifies later on that the level of analysis is the grade rather than the teacher, but his clarification is far more technical: “So let’s say there are four teachers teaching in a particular grade, and one of them is a highly ineffective teacher who leaves, we’re not saying look at the test scores of that one teacher. We’re saying let’s average the test scores of all the fourth-graders. Now, 25 percent of them were exposed to the highly ineffective teacher. Now we’re going to compare that to all fourth graders’ test scores on average after that teacher leaves, and the average teacher quality that those students face in the next year will be higher…the idea is that on average the entire grade has better teachers and that’s the nature of the experiment.” (pg.500)
The next part of Chetty’s testimony focuses on forecast bias. This concept [described in footnote 23 above] is very sophisticated and relies on a complex estimation procedure, which the witness and examiner make no attempt to describe. The demonstrative evidence is the binned scatterplot charts presented in the *Long-term Impacts* paper, which nicely show points clustering around a best fit line, but compress nearly 60,000 data points into only twenty, giving a false sense of consistency (slides 214-01 and 214-02). The reported slope of the best fit line (0.98) is very near a perfect correlation, and looks more perfect because the outliers among the 60,000 points are compressed into the binned average points.

The testimony then turns to the actual long-term impacts cited in the paper, beginning with college attendance and earnings. A major omission of the testimony is the way “long-term impacts” are actually determined. The authors infer the long-term impacts of teachers based on the correlations between test scores and later outcomes. This is at the heart of Ballou’s (2012) critique of their study discussed above. To know that teachers have such long-term impacts, one first needs to determine non-spuriousness – whether other variables which correlate with teacher value-added are also correlated with long-term outcomes.

Chetty’s exhibits include three charts taken from the *Long-Term Impacts* paper, which are again the binned scatterplots – though this time 20 data points stand in for 1.1 million student observations, which Chetty points out in his testimony (p. 515). This means that 55,000 student observations are represented by each of these points.

---

195 “So let me describe how we construct it. On the x-axis are changes in average value-added of teachers across consecutive cohorts of students. So like we were using in the [quasi-experimental] example before, this would be the difference in the average value-added of teachers in 1996 minus 1995 in a given school in a given grade in a given subject, okay. And on the y-axis is the change in the actual test scores for the students between 1996 and 1995 or between the next year and the previous year generally.” This description omits the detail that average teacher value-added for a grade-subject in ‘1996 minus 1995’ is estimated without data from the years 1996 and 1995, and further likely includes data from years 1997 and later.
data points are being compressed into one (slide 204-01). Further, even though Chetty’s testimony mentions highly effective and highly ineffective teachers (the top and bottom 5 percent of the distribution); the lowest bin on the chart’s x-axis indicates an average ranking in the 15th percentile, while the highest point is at the 90th percentile.\textsuperscript{196} The chart’s y-axis is equally important to consider, as the range of teacher impacts on college attendance is less than 2 percentage points. Chetty argues that this represents the impact of a single teacher, and that the effects of teachers are cumulative, arguing that the college attendance effect could be “10 or 15 percentage points,” calling the potential effects “roughly additive.” (p.517). Value-added research – including Kane and Staiger’s Los Angeles experiment – has found that teacher effects fade out drastically (Kane & Staiger, 2008; Jacob, Lefgren, & Sims, 2010). Nonetheless, Chetty argues that this cumulative benefit can also result in cumulative harm in the case of successive ineffective teachers. This point about cumulative teacher effects is reiterated in his testimony regarding college quality and student income.

The discussion of teachers’ impacts on students’ college quality begins by establishing the measure used. Chetty and his coauthors gauged a college’s quality based on the mean earnings of its graduates 12 years later.\textsuperscript{197} Chetty refers to this measure as simultaneously indicating a

\textsuperscript{196} The chart illustrates a point made in Chapter 2; value-added rankings are not normally distributed – teacher VA scores tend to cluster in the middle of the distribution. Indeed, five binned data points are clustered around to the 50th percentile. Yet Chetty goes on “You can directly see from this chart real data with these 1.1 million children. A child who is assigned to a teacher who’s in the bottom 5 percent of the distribution, a highly ineffective teacher, based on value-added metrics, has a 36 percent chance of attending college in this school district compared with approximately 38 percent for a child who is assigned a highly effective teacher in the top 5 percent of the distribution.” (pg. 516)

\textsuperscript{197} At this point in the testimony, Chetty indicates his involvement with the Obama administration’s controversial attempt to rank colleges based on student’s mean post-graduation earnings. The initial plan, which would have tied college funding to these ratings, ultimately failed in Congress, but the Department of Education’s “college scorecards “ still do report this statistic (Fain, 2013; Stratford, 2015).
college’s selectivity and its value-added (p. 521). He argues that a single highly effective teacher can increase her students’ average college quality by two to three percentage points or $500 in additional average post-graduation earnings when compared to a highly ineffective teacher. The eighteen points in the middle of the chart are not discussed, as the differences between their effects on college quality are far less dramatic than two percentage points.

Chetty’s testimony then focuses on the earnings impacts of highly effective and highly ineffective teachers, using two charts. The first chart (slide 205-01), drawn from Long-Term Impacts, focuses on students’ reported earnings at age 28, shows far more variation than those related to college outcomes. As such, the best fit line looks flatter than in the three previous figures – the points associated with the two highest ventiles are no higher than some near the median. Chetty explains this as statistical noise – despite the analysis being based on over 360,000 observations – and instead directs attention to the trend line.\footnote{He does concede that it is possible that “teachers in the very upper tail of the distribution, the top 1 percent or the top 2 percent of the value-added distribution might not be having as substantial effects on students’ long term earnings as you might predict based on their impacts on test scores.” This point tacitly reveals the methodology of the Long Term Impacts study, which was described above, but not discussed explicitly in the testimony.} It is of interest that he does not cite the $182 difference between the bottom and top five percentiles, which is perhaps the most cited figure from the paper. He is saving dollar amounts for a larger idea.

The next chart deviates from those included in Long-Term Impacts. It presents a normal curve with the bottom 5-percent highlighted. He proposes that if schools were to ‘let go’ of teachers in this lowest value-added ventile and replace them with average teachers, that the data suggest this would improve the lifetime earnings of each child by $50,000, and multiplies this figure by 28 – the average class size in the data – to produce a classroom-level effect of $1.4 million. The
$50,000 figure is said to represent a 2 to 3 percentage point increase in lifetime earnings, and is further exploited by comparing it to the impact of the “Great Recession” which reduced GDP by between 6 and 7 percent. Thus he argues that the “feasible policy” of removing and replacing low value-added teachers with average ones would have an effect “on the same order as ending the financial crisis again and again, year after year.” (p. 529)

This point needs detailed explication. As noted earlier, these effect estimates assume little fade out of teacher effects over time – a fourth-grader is nine years old and the testimony discusses income measured at age 28. Further, the testimony did not include the authors’ measures of income effects at age 30, which were not significantly different from zero. Even the paper did not examine income at other age points under 30 where participation in the labor force is robust. This “feasible” policy of effectiveness-based layoffs was first mentioned by Eric Hanushek in the propaganda film Waiting For Superman. But for such a policy to be feasible, school systems would need to have a reserve army at least five percent of the size of their teacher labor forces available every year and adequately divided by grade and subject specializations.

Chetty’s testimony then briefly moves through three other outcomes examined in Long-Term Impacts: teen pregnancy, neighborhood quality, and retirement savings (slides 206-01 through 206-03). He argues that being assigned to highly ineffective teachers increases the likelihood that

199 Chetty discussed fade-out briefly in the testimony. He acknowledges the fade out of test scores documented in the value-added literature and argues that this is why they chose to study long-term impacts directly (pg. 537). But this holds only if one accepts studying teachers’ effects on earnings via her average marginal effect on test score growth as a direct measure.

200 For example, the most recent Baccalaureate and Beyond (B&B) survey indicates that about 80 percent of recent bachelor’s degree recipients were employed by age 26 and that a large majority of these held full-time jobs (Cataldi, Siegel, Shepherd, & Cooney, 2014). Labor force participation among two-year college students and high-school graduates who did not attend college is probably comparable. This suggests that Long Term Impacts could have investigated at least three other age points with large sample sizes.
female students will have children at or before age 19, reduces the quality of the neighborhood that students live in at age 28, and negatively impacts the probability of students having retirement savings accounts by age 28. (pgs. 529-532) At this point, Chetty again quickly reveals the arbitrary analytic choices made in the study – focusing their analysis on outcomes for 28 year-olds. Chetty says that this choice was made “simply because we don’t have an adequate number of children who are beyond age 28 in our data.” (pg. 532) As Adler (2013) noted in his critique of this study, this inadequate number of students is over 61,000 – quite a robust figure by most standards. The next slide (205-02) returns to the earnings outcome and tracks the authors’ estimated impacts over time from ages 20 to 28. Chetty uses this slide to show that teacher effects on student’s earnings trend upward.\textsuperscript{201}

The next set of comments in Chetty’s testimony respond to three criticisms of value-added measures made by other researchers: that VA cannot disentangle the many factors which influence student test scores, that VA scores are affected by the students to which teachers are assigned, and that VA measures fluctuate from year to year. To the first point, he begins by agreeing that many things can affect student test scores; he then makes an analogy to basketball to argue that even though many factors affect performance, policy should try to control what it can.\textsuperscript{202} To the second point, Chetty argues that his teacher-switching quasi experiment as well as

\textsuperscript{201} But this slide also reveals the authors’ selective presentation of their own data. The chart ends at age 28, but includes all prior data points. This is inconsistent with the paper, which presented earnings estimates for age 30, but for no other age points. The y-axis is the percentage point impact on earnings of a teacher one standard deviation above average. The authors correctly explain the negative impacts before age 23 – many students are still in college (testimony pg. 533). After that, we see that the largest single-year jump in the earnings effect occurs between age 27 and 28, and that between ages 24 and 26, the average effect is about 1 percentage point. One possible explanation for not including these findings in the paper or the testimony is that these effects did not achieve statistical significance.

\textsuperscript{202} “There are many determinants of a basketball player’s ability. One of the most important is perhaps a player’s height. If you are 5’2”, your odds of making it to the NBA are not great
Kane and Staiger’s random assignment study in Los Angeles have proved that value-added measures are unbiased if they control for prior test scores. Though many researchers agree that the *Long-Term Impacts* paper showed that value-added was capable of capturing teacher contributions to test scores, some have still raised objections to their technique of estimating bias (Rothstein J., 2015). Further, reviews of the random assignment study in Los Angeles showed that important groups of students were likely excluded. To the final point, the unreliability of VA measures, Chetty later responds that simply finding a low correlation between value-added relative to if you are 6’6” or 7 feet tall…Now does this mean that you shouldn’t be concerned about coaching a basketball player, like should we be interested in trying to improve coaching? My view is the fact that there are other important determinants of a player’s ability doesn’t mean that you should not pay any attention to things you can manipulate.” (p. 540) This is a weak defense of value-added as a policy tool; as the proper analogy would be if some coaches got teams full of 5’2” players while others got teams full of 6’6” players and we nonetheless had a policy of firing those coaches with the lowest average point totals. Chetty makes extended analogies to baseball and basketball statistics to support numerous points in his testimony, so it seems appropriate to unpack the limits of such an analogy. There is a deep affinity between value-added assessment in education and statistics in sports management. Baseball analysts have long kept track of certain basic statistics – runs scored, batting average, bases stolen, earned run average, etc. But beginning in the 1980s, some baseball researchers began applying the tools of statistical inference to “search for objective knowledge about baseball” (Grabiner, 1994). The resulting approach – known as Sabermetrics – created new statistics for the analysis of the sport which measure players’ contributions accounting for their positions and other aspects of their role on a team. The most famous application of these principles was on the Oakland Athletics professional team in the 1990s, which was documented by Michael Lewis in the book *Moneyball*. One statistic – value over replacement player (VORP) – is precisely the same operative idea behind value-added. A replacement player is defined by certain average characteristics for backup players, and active players’ contributions to their teams are quantified by their deviation from this average (Woolner, 2001). In baseball there are acceptably talented players far in excess of the available positions on major league teams (1,200 roster positions are available in all). In such a situation, it makes sense to create management policy based on the assumption that average players are available to replace your starting players – especially at a league minimum salary of over $500,000 per year. However, with a teaching labor force of 3.1 million where salaries average $55,000 per year, a value over replacement teacher strategy is hardly tenable. Further, baseball has only one “important” outcome – winning games. Nonetheless, VORP combines numerous statistics to measure how players contribute to that goal. Education reformers instead seek to use one measurement – student test scores – to assess teacher contributions to an undetermined number of socially important outcomes.
scores from one year to the next is not enough to render the data invalid. He instead returns to his dollar-value calculations of the impact of using value-added as a policy tool and says that, regardless of how accurate the measure is, an effectiveness-based layoff policy would produce $1.4 million in additional income for students, so it is worth it.\textsuperscript{203} His response to this criticism also relies on sports analogies; he notes that batting averages are unreliable and that despite basketball players’ average performances, people tend to focus on outliers (p. 554). So with the exception of the question of forecast bias, for which his study provides an answer, Chetty’s response to criticisms of VA measures is that whatever their flaws, they should be put into practice. Broadly, he hinges his belief in the validity of value-added methods on the teacher switching quasi-experiment featured in Long Term Impacts. We have seen that this is a strong but not infallible piece of evidence. It also does not by itself prove teachers’ long term impacts or confirm value-added as standalone measure of teacher quality.

The plaintiffs finally ask Chetty to comment on two LAUSD policies – a 16-month teacher probation period and the “Last-in-First-out” (LIFO) teacher retention policy. Chetty’s comments on these policies both use VA methods as a reference point and invoke to the theorized policy of firing teachers in the bottom five percent of the effectiveness along with the hypothesized lifetime income benefits it would afford to students. He argues that the probationary period should be 3 years instead of 16 months, since the additional years of value-added data would more precisely identify the bottom of the distribution for more finely-tuned teacher layoffs – though he uses evasive terms “deselecting the bottom five percent” (p. 567).\textsuperscript{204} Lengthening the

\textsuperscript{203} “…and that’s the critical point here. The measure is not perfect. Qualitatively it’s correct that there is inconsistency, but, that, in my view, a 1.4 million-dollar gain is still very substantial and does not undermine the value of the measures.” (p. 553)

\textsuperscript{204} The term “deselecting” was first used to describe this policy by Eric Hanushek (2009).
probationary period, he argues, would increase students’ lifetime earnings by $163,000 (Slide 208-03).

Chetty’s comments on LIFO make use of data analysis commissioned by the plaintiffs rather than *Long-Term Impacts* – thus it is based on analysis of LAUSD data which are significantly smaller (500,000 students over three years from 2009-2011) than the New York City data (2.5 million students over twenty years). The claims he makes about the impacts of different policies also require many assumptions about school policy. For example, the court is asked to imagine that Los Angeles is asked to lay off five percent of its teacher labor force (p. 573). This is done to contrast the possibility of laying off teachers based on effectiveness ratings against doing so based on seniority. He notes the LIFO policy results in layoffs across the value-added distribution, including many ‘above average’ teachers and some in the top five percent of the distribution. Of the latter group of high value-added teachers who would be affected, Chetty argues that “we should be doing everything we can to keep these superstars of the teaching profession in our school district.” (p. 575) Again returning to sports analogies and says that seniority-based layoffs would be akin to cutting Michael Jordan from the basketball team if he didn’t do well in his first year. This is dubious; an effectiveness-based layoff policy would be as likely (if not more so) to cut a player who didn’t do well. Here his argument against LIFO complicates his position on extending the probationary period. If probationary periods – during which teachers can be fired with little recourse – were extended to three years, many new teachers might be “deselected” before their “superstar” status could be determined.

---

205 Chetty testifies that about “roughly five percent” of the teachers who would be laid off under a LIFO policy would be in the top five percent of the value-added distribution (p. 575). But if his demonstrative chart of two largely overlapping normal distributions (slide 201-3) is to be believed, that percentage is substantially smaller.
Chetty goes on to describe the earnings impacts of switching to a layoff policy based on the VA distribution. This combines the analysis of test score data in Los Angeles with the earnings impacts estimated in New York City – which by fiat makes the assumption that such impacts can be inferred from one municipality to another. He argues that the impact of removing teachers in the bottom five percent of the VA distribution would increase students’ lifetime earnings by $2.1 million for every teacher laid off (p. 579). After briefly addressing the disproportionate effects of LIFO policies on poor and minority students – based on Thomas Kane’s research – Boutros ends his examination of Dr. Chetty.\footnote{206}

Expert Witness Testimony in Vergara – Thomas Kane

Dr. Thomas Kane testified over two days from February 5\textsuperscript{th} to February 6\textsuperscript{th} 2014.\footnote{207} He was examined by He states that his primary research interest since 2006 has been teacher effectiveness, and that from 2008 to 2012 he was on leave from Harvard to work as deputy director of K-12 programs at the Bill and Melinda Gates Foundation (Feb 5 PM Session,

\begin{footnotesize}

206 The cross-examination is shorter and far less focused, so I only discuss it briefly. James Finberg, the first cross-examining attorney for the defense, begins by confirming that Chetty has never actually taught or been an administrator in a K-12 classroom (p. 587). But he seemed too unfamiliar with VA methods or with Chetty’s testimony to effectively cross-examine. After failed attempts to talk about Chetty’s Long-Term Impacts study, Finberg questions Chetty on in-school and out-of-school factors other than teacher effectiveness that could impact student achievement, establishing from Chetty’s research that smaller class size and teacher experience were also shown to be correlated with better long-term student outcomes (p. 595-602). He further confirms from Chetty’s own research that cities in California including Los Angeles were ranked high on a social mobility index, and that the quality of local schools was one important factor in determining levels of social mobility (p. 606-609). Nimrod Elliot, the second cross-examiner, confirms Chetty’s hourly retainer rate as an expert witness, $825 per hour and that the portions of his expert testimony on LIFO policies were conducted as an expert witness and not subject to peer review (p. 610). His final questions clarify some of Chetty’s model specifications on the LIFO analyses.

207 Since there were no readily available transcripts, I rely on a video of Kane’s testimony from the Courtroom View Network. References refer to the date and time of the session and the time stamp of the video. Citations to the slides are the same as done with Chetty’s testimony.

\end{footnotesize}
208 It was during this time that he worked on the MET project with the specific goal of developing “systematic, replicable measures of teacher effectiveness.” (Feb 5 PM Session, 2h47m).

The first remarks Kane makes with respect to teacher evaluation concern standard classroom observations, which he regards as “perfunctory.” (Feb 5, PM, 2h48m) He then gives the main point from the five MET study reports – that they have identified stable, unbiased measures of teacher effectiveness and that those measures correlate with observed student achievement gains (Feb 6, AM 0h1m). He mentions a number of his teacher effectiveness studies, including a replication of Chetty’s Long Term Impacts (Bacher-Hicks, Kane, & Staiger, 2014) and the random assignment study in Los Angeles which he referred to as a miniature MET project (Kane & Staiger, 2008) (Feb 6, AM, 0h7m).

Kane was retained to comment on five topics, three of which are conceptual in relation to teacher effectiveness, and two which are specific to Los Angeles schools (Slide 356-01). His opinion on Los Angeles is that the district shows wider variation in teacher effectiveness than do many other cities and that Black and Hispanic students are disproportionately assigned to ineffective teachers. But he begins by discussing teacher effectiveness generally and the MET study in particular.

208 Kane connects his selection to head the MET study to his work with Doug Staiger and Robert Gordon, “Identifying Effective Teachers Using Performance on the Job” (Gordon, Kane, & Staiger, 2006).
209 He describes systematic and replicable measures later on in his testimony and clarifies that this refers to student surveys and scored teacher observations in addition to student test scores (Feb 6 AM 0h21m). Kane generally seems more careful in his comments about Value-Added models.
The first few minutes of testimony are primarily drawn from a Gates Foundation Policy and Practice Brief on *Ensuring Fair and Reliable Measures of Effective Teaching*. Kane starts his opinion by saying that is essential for school districts to measure teacher effectiveness. He supports this by saying that measurement is essential so that teachers can “develop their own effectiveness” through feedback, and that not doing so would be like “running Weight Watchers without bathroom scales or mirrors” (Feb 6, AM 0h18m). He advocates direct measures of student achievement gains where possible and other data – observations and student surveys – which are known to be related to student achievement gains (Slide 359-01). He distinguishes between these measures and others – such as teacher credentials and certifications – which have been shown to be uncorrelated with student test score gains (Feb 6 AM, 0h26m).

Kane and Lipschutz (the plaintiff’s attorney) engage in a discussion of Campbell’s Law (Slide 360-01). The quote from *Assessing the Impact of Planned Social Change* refers both to the

---

210 Rothstein and Mathis (2013) point out that these policy briefs do not provide the level of detail of the full papers and give the impression that random assignment was unproblematic.  
211 The slide quotes “The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor…Many commentators, including myself (1969b), assume that the use of multiple measures, all recognized as imperfect, will alleviate the problem, although Ridgeway (1956) doubts this.”

The quotation is misleading. The ellipsis – which is intended for use when the meaning of a quote is unchanged by the omission – omits ten pages of text. In those ten pages, Campbell discusses at length the “highly corruptible” nature of student achievement tests (Campbell, 1976, p. 61), the phenomenon known as “creaming” – where the most difficult cases are excluded from accountability metrics (p. 60), and the need to have practitioner organizations cooperate in the measurement process using the example of a teachers’ union (p. 63).

In the same paper, Campbell explicitly urges evaluation researchers to refuse to conduct the sort of research that Kane and Chetty have advanced. His point is worth quoting at length – and without substantive omission: “I recommend that we evaluation researcher methodologists should refuse to use our skills in *ad hominem* research. While the expensive machinery of social experimentation can be used to evaluate persons, it should not be. Such results are of very limited generalizability… (5 lines down) This prohibition on *ad hominem* research should also
corruption of quantitative indicators and to Campbell’s support for using multiple measures – but the attorney pushes Kane’s testimony to focus on the value of value-added as a single quantitative indicator (Feb 6 AM, 0h36m).

Nonetheless, Kane makes an interesting point here; triangulation of multiple sources of data around a concept is an important way to validate findings, and perhaps to improve evaluation policy. But one also needs to question how much the MET project really promoted ‘multiple measures.’ As we saw above, the measure of an important measure is, for MET, its correlation with Value-Added. Kane emphasizes this both in the MET reports and in his expert testimony. So if all the measures are conceptually bound to a single measure, how many measures do we really have?

The testimony then turns to the MET findings on multiple measures – Value-Added, Student Surveys, and Classroom Observations. Kane’s remarks here center on whether the combined metrics developed in the MET project perform better than existing methods of making teacher tenure and retention decisions. Similar to other proponents of VA, Kane evaluates value-added be extended to program clients. We should not be evaluating students or welfare recipients but alternative policies for dealing with their problems.” (Campbell, 1976, p. 56)

212 The exchange is important. “Now does Campbell’s law mean that you cannot measure teacher effectiveness using Value Added Measures? No. Why Not? The implication of the first statement is that if you used just Value-Added, there’s a danger that it would become distorted. But the second part of Campbell’s Law is implying that the way to alleviate that problem is to collect multiple measures that would allow you to lessen the chance that teachers or principals or whomever would improperly try to distort the behavior you’re trying to incentivize. But even if you used only Value-Added measures, would that provide you any insight into a teacher’s effectiveness? Yes. If one only had a Value-Added measure, one could characterize differences in effectiveness, but one wouldn’t want to use that as the sole measure and apply it in an evaluation system because the measure may no longer be measuring the things that you think it is. But if you take an existing data set and apply Value-Added to that existing data set where Value-Added is not used for high-stakes purposes, can you use Value-Added to assess teacher effectiveness? Yes.”
not on its own strength, but on its comparative merits (see for example (Harris D. N., Value-Added Measures in Education: What Every Educator Needs to Know, 2011)). The chart which informs this testimony demonstrates a circular logic; existing measures of teacher quality are compared with the MET composite based on their respective correlations with student test score gains – one component of the MET composite. This same circular logic is evident in one of MET’s three culminating reports (Mihaly, McCaffrey, Staiger, & Lockwood, 2013), where the authors find that each measure of effective teaching is only strongly correlated with itself.

On the same point (Slide 365-01; Slide 366-01) Kane begins to use a misleading measure of student achievement to support his position on the use of VA measures. His slides convert standardized test scores to months of student learning (Feb 6 AM, 0h50m-56m). Later in his testimony, one chart indicates that this procedure of converting test score gains to months follows Hill et al. 2008 (Slide371-01). That paper – originally released as a technical report released by the research firm MDRC – estimates average annual student test score growth by grade level; it never endorses the practice of converting test score gains into months of learning (Hill, Bloom, Black, & Lipsey, 2007).\(^{213}\) As noted above, this conversion was first used by Hanushek (1992) and has been subject to criticism.

The next section of Kane’s testimony focuses on the MET random assignment study. As with Chetty’s testimony, random assignment is described as the only way to make causal claims based on research (Feb 6 AM, 1h05m). He then relies on slides which plots predicted and observed

\(^{213}\) Specifically, Hill et al. includes a table which shows the average student test score gains by grade level across six nationally-normed math tests and seven nationally-normed reading tests (p. 3). Because the MET study focuses on grades 4 through 8, Kane averages the annual test score gains across those five grade levels (.420 standard deviations) and then divides by twelve to calculate the average month of student learning – .035 standard deviations (author’s calculation based on Hill et al. 2007 and Kane et al. 2013).
student achievement gains (Slide 369-01, 370-01). Unlike the previous slides, predicted teacher effectiveness is only measured by value-added. The chart is a binned scatterplot like those featured in Chetty’s testimony; in this case, each of the twenty points represents 40 teachers, meaning that the plot could have instead featured 800 points (Feb 6 AM, 1h10m-13m). This is different from Chetty’s charts insofar as it would have been feasible to plot 800 points without too much clutter; however, it would likely have been a far less persuasive picture. In both slides (Math and ELA), the slopes of the regression lines are much more shallow than those in Chetty’s *Long Term Impacts* papers. Kane ultimately states his opinion based on this evidence: that teacher probationary periods should be extended. He does this relying on an odd analogy to dating. Contrary to Kane’s emphasis on multiple measures, his discussion of the benefits of extending the probationary period for new teachers focuses solely on the additional years of test score data which that extension affords (Feb 6 AM, 1h38m).

Kane’s testimony on his second topic – differences in teacher effectiveness – relies on his study of teachers in the Los Angeles Unified School District (LAUSD), which replicates the first section of *Chetty et al.* and relies solely on VA measures of effectiveness (Feb 6 AM 1h56m). He defends this study against his emphasis on multiple measures, arguing that the point of the LAUSD study was to validate VA measures, not to recommend VA as a standalone teacher evaluation policy. It is important to note that in the paper based on these data, Kane and his 214 “Now do you have an opinion about whether a year and a half is enough time for a school district to gather the evidence…to make an informed judgment about a teacher’s effectiveness? My opinion is that it’s not long enough; it’s analogous to telling people that they can have a maximum of one-and-a-half dates before they decide to get married. Certainly after one date it’s possible to identify matches that won’t work out, but I don’t think it’s possible after one-and-a-half dates to be really confident that this commitment you’re making is worthwhile. If we limited people to one-and-a-half dates before making marriage decisions, we’d have a lot of bad marriages” (Feb 6 AM, 1h34-36m). This analogy is simply too absurd to seriously consider, but Kane insists on this logic, returning to it after being cross-examined (Feb 6 PM, 1h48m).
colleagues do not mention the need for multiple measures as they justify the validity of VA measures as a policy tool (Bacher-Hicks, Kane, & Staiger, 2014). The testimony notes three findings from the LAUSD study: that teacher effectiveness measures are predictable, that the effects of teacher assignment are larger than those found in New York, and that ineffective teachers are disproportionately assigned to Black and Latino students (Slide 364-01). The first finding is supported by a binned regression plot similar to those used in Chetty’s testimony (Slide 351-01, cf. Slide 214-01) and similar statements by Kane (Feb 6 AM, 2h00m).

The argument regarding the effects of teacher assignment on learning is supported by charts comparing the effect size differences in the LAUSD study to those in Long Term Impacts and the MET study (Slides 371-01 and 372-01). Specifically, Kane notes the learning differences between the bottom 5 percent and average teachers, using the same months of learning conversion described above, across the three studies. This makes the effect of teacher assignment in Los Angeles look far more dramatic than in other municipalities; the chart shows students with the lowest VA teachers losing 9.5 “months” of ELA learning and 11.7 “months” of math learning – an entire school year. This is compared to 6.6/7.0 months in the New York (as found by Chetty et al.) and 4.8/6.7 months in the six MET districts. The prosecutor emphasizes this comparison with the following exchange:

“That means that I lose 11.73 months or an entire year of learning relative to students assigned to the average teacher. How is that possible, Dr. Kane? How could a student lose an entire year of learning in one year? By not learning very much…So what is the conclusion we should take from this slide? Although there are differences in all three districts associated with being assigned to a bottom fifth percentile teacher rather than the average teacher, the differences are larger and the consequences are larger in Los Angeles…” (Feb 6 AM, 2h06m)

Kane makes two final points regarding the LAUSD study. The first is that Black and Latino students are assigned to teachers who have on average lower VA scores. The second is that those
students are more likely to be assigned to teachers in the bottom five percent of the VA distribution. His demonstrative slide again uses the months of learning metric. Kane refines the point by suggesting that the “maldistribution” of teachers in Los Angeles is widening racial achievement gaps. But in making this point he also notes that students in the data who started a year with lower test scores were more likely to be assigned to low VA teachers.\footnote{“We also looked at differences in the effectiveness of teachers assigned to students who started the year low on state achievement tests to those who started the year high on state achievement tests. And the relationship we saw was that students who started the year lower on the state tests were assigned less effective teachers one average than [sic] students who started the year with higher achievement on state tests.” (Feb 6 AM, 2h25m) In other words, the best teachers are assigned to the best students – one wonders how these teachers got to be the “best.”}

Inadvertently, this reinforces arguments for student sorting made by Rothstein (2010) and others (Garrett & Steinberg, 2015). In describing the reasons for the maldistribution of teachers, he says that policies which make premature tenure decisions or make it hard to fire ineffective teachers could lead to what he calls a “Lemon Accumulation Machine.”\footnote{He describes this phrase in more detail. “In any system where districts have to make tenure decisions prematurely and where it is difficult to make dismissal decisions later, ineffective teachers will tend to accumulate. It’s sort of an inevitable result of those two factors and the regrettable implication is that ineffective teachers then tend to accumulate in the schools where there are more vacancies. And the schools where there are more vacancies are often the schools where there are large accumulations of African American or Latino students.” (Feb 6 PM, 0h1m) One gets the impression from this remark that the same forces that make a school desirable to teachers are the same environmental factors that lead to higher test scores – and consequently to higher teacher VA ratings. I demonstrated this in Chapter 2.} He further says that low VA teachers tend to settle in schools with higher proportions of black and Latino students owing to labor market forces (Feb 6 AM, 2h26m-28m).

The direct examination concludes after Kane briefly addresses how the MET study complements the findings of \textit{Long Term Impacts}. James The first cross-examination on behalf of the California Teachers’ Association is conducted by James Finberg. He begins in the same way as with Chetty, confirming that Kane has never been a teacher or administrator in a public school. But
his cross-examination of Kane is much more substantial, lasting over an hour. As with the direct examination, I only focus on portions which pertain to VA assessment.

The defense attorney first establishes, based on language from one MET study (The Bill and Melinda Gates Foundation, 2013), that Kane prefers evaluations which use multiple measures, and that such evaluations produce more consistent estimates of teacher effectiveness. He then confirms Kane’s belief that relying on a single indicator (such as state test scores) can have negative consequences for students, teachers and schools (Feb 6 PM, 0h15m-35m).²¹⁷ Relying from the LAUSD study, the defense establishes that teacher mal-distribution was related not only on race/ethnicity but also to prior-year student test scores, both within and between schools (Feb 6 PM, 0h38m-44m). Using Kane’s Hamilton project study (Gordon, Kane, & Staiger, Identifying Effective Teachers Using Performance on the Job, 2006) he establishes that Kane believes that working conditions affect teachers’ labor market decisions, that schools can get information about teacher effectiveness using less than three years of data, and that the authors did not advocate dismantling or substantially altering the teacher tenure system (Feb 6 PM, 0h50m-1h11m). He finally utilizes the work of other scholars (Rothstein J., 2010; Papay & Kraft, Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement, 2015) to confirm with Kane that test score

²¹⁷ Across all the witnesses I considered, cross-examinations were labored by objections from the plaintiffs. In Kane’s case, it took nearly 20 minutes to establish these two points. The defense attempted to establish that Kane doesn’t believe multiple measures are necessary to identify effective teaching, but the plaintiffs’ objection to this line of questioning was sustained by Judge Treu. Generally, the judge seemed very impatient with the defense’s questioning method.
impacts fade over time and that teacher experience is positively related to student achievement gains (Feb 6 PM, 1h12m-1h15m).\textsuperscript{218}

Kane is also briefly examined by an attorney for the State of California. She establishes that Kane believes multiple measures to be important to prevent ‘gaming’ of the tests or cheating, and that such would be a consequence of sole reliance on VA measures (Feb 6 PM, 1h17m). She also establishes that the study did not consider whether individual teachers were teaching out-of-field, which she calls “mis-assignment” (1h20m). She finally establishes with Kane many in-school and out-of-school factors other than teachers affect student performance (1h24m).

Expert Testimony – Jesse Rothstein

Attorneys for the defense called UC-Berkeley professor Jesse Rothstein as an expert witness on March 6\textsuperscript{th}, 2014. Direct examination lasted four hours; cross examination took an additional hour, making Rothstein’s the longest testimony among the experts on value-added. The length of his testimony owes both to the frequency of procedural objections made by the plaintiffs’ legal team, and to the fact that Rothstein addresses both value-added models and the broader issue teacher tenure and dismissal policies.

The substantive basis for the first part of Rothstein’s testimony is his paper “Teacher Quality Policy When Supply Matters” (2012). The paper responds to the hypothetical policy of annually

\textsuperscript{218} Papay and Kraft’s (2015) study was originally released as a working paper in 2014, and concludes based on North Carolina data that teachers continue to grow in effectiveness (as measured by Value-Added) beyond the first three years. This calls into question the prevailing wisdom from economists that VA levels off after three years – which is the rationale for the probationary period proposed in Chetty and Kane’s testimonies. In a related study, they also find that these returns to experience vary based on the level of support teachers receive in their schools (Kraft & Papay, 2014). This again suggests that context matters, a point largely elided by VA proponents.
firing teachers in the bottom five percent of the effectiveness distribution and to that of prolonging the mandatory teacher probation period by posing the issue of teacher supply. Given that no municipalities have yet implemented such policies, Rothstein uses simulations to infer how such policies would alter the teacher labor market. The idea of simulation research initially confuses the court (Mar 6 AM, 0h23-25m) and prompts numerous objections from the plaintiffs (Mar 6 AM, 0h23m-43m).²¹⁹ In this portion of his testimony, Rothstein addresses a number of broad concerns about high-stakes teacher evaluation policies in terms of new and continuing teachers’ perceptions of their job security.

Following the arguments made in his paper, Rothstein suggests that potential teachers consider a number of factors when deciding to enter the field. Since the pay is low relative to other jobs for college graduates, teaching has to appeal to new recruits for non-salary reasons – job security being important among them. Further, since teachers’ earnings tend to be “back-loaded” – greater at the end of their careers – job security remains an important consideration among continuing teachers (Mar 6 PM, 0h1m-3m). Given this, Rothstein contends that evaluation systems and tenure policies which increase the likelihood of arbitrary or capricious layoffs serve to reduce the attractiveness of the profession both to potential new teachers and to existing teachers, and make it harder to recruit and retain effective educators (Mar 6 PM, 0h18m). His early testimony only tangentially addresses value-added models. For example, he argues that the economic returns to extending the probationary period suggested by Chetty in his testimony – and based on the Long Term Impacts paper – would be offset by the broader teacher labor market

²¹⁹ Chetty’s evidence on teacher dismissal policy, though based on empirical data both rely on simulated counterfactual policies. Many early studies of VA methods similarly relied on simulations like those conducted by Rothstein, including Hanushek’s early work. However, their conjectures prompt no objections from the defense. As we have seen, this seems to be a difference of style between the two legal teams.
effects of such a policy (Mar 6 AM, 1h34m-35m).\textsuperscript{220} He then makes a case that employing VA instead of seniority-based rules in reductions in force would negatively impact teaching in a number of ways including increased cost of evaluations, and decreased attachment of teachers to their schools and communities (Mar 6 PM, 0h24m-33m). In short, this section of Rothstein’s testimony suggests that using effectiveness-based layoff policies could do substantive harm to schools by placing too much emphasis on indicators taken out-of-context. Ironically, teacher dismissal policies that explicitly seek to improve aggregate teacher effectiveness could end up having the opposite effect if they in the process make teachers less willing to enter or remain in teaching.

Rothstein’s testimony then shifts to value-added models and the testimony given by Raj Chetty and Thomas Kane. In addition to echoing some of the caveats presented above, Rothstein’s critique of Chetty presents two additional problems which compromise his claim that VA measures are unbiased. First, Chetty’s concept of bias is shown to fundamentally different from what other VA researchers. Whereas most research on value-added holds teacher ability constant when testing for bias, the bias tests featured in \textit{Long Term Impacts} allow teacher effectiveness to grow or diminish over time. According to Rothstein’s testimony, Chetty Freidman and Rockoff changed this specification between versions of their study, and essentially “redefined bias as non-bias.”\textsuperscript{221} Second, he argues that Chetty et al.’s teacher-switching experiment – which was

\textsuperscript{220} Rothstein also argues that a lengthened probationary period proposed by VA proponents would harm student achievement because ineffective teachers would likely remain in the system until they were denied tenure (Mar 6 AM, 1h10m-12m).

\textsuperscript{221} “And in the first version of the study, he found that it didn't, so he found what I would interpret as evidence for bias. In the second version of the study based on feedback that I gave him and I'm sure that others gave him as well, there was a new interpretation of that evidence, and so he changed the question somewhat. So in the past, all VA studies assume that teachers have a fixed level of effectiveness, and that what we were doing with VA was trying to measure that; but that the teachers underlying level of effectiveness didn't change from year to year. And
also used to prove non-bias – was not a very strong test. While the paper implies that teacher-switching only occurred between schools, many of the actual incidents of switching were grade changes within schools (Mar 6 PM, 0h43-45m). This is a consequential difference, since VA research (including Chapter 2 above) shows that teacher effectiveness estimates are affected by school context. Teacher-switching within schools therefore doesn’t demonstrate the robustness of teacher quality, and serves to contaminate Long Term Impacts’ evidence of non-bias.

The defense then asks about the strength of Kane’s random assignment studies as evidence of non-biased VA estimates. Again, Rothstein offers two additions to the caveats already detailed above. He notes that the mini-MET study in Los Angeles (Kane & Staiger, 2008) was a voluntary opt-in study and that the data show the teacher pairs selected by principals to be very similar:

“Also in that study the pairs of teachers he had were carefully selected in the sense that principals were allowed to nominate teachers to participate in it, and many of the principals who they asked to participate declined to participate, and so one of the concerns is that the pairs of the teachers they were looking at in that study were more similar to each other than our normal pairs of teachers. They're teachers for whom the principal doesn't care which students are assigned to which teacher. And so that's a setting where I would expect there to be relatively little bias, and it doesn't tell you much about the broader setting.” (Mar 6 PM, 0h48m)

Although Rothstein acknowledges that although the MET study was far larger and therefore “less selected,” he contends that the non-compliance was systematic in ways which would tend to reduce estimates of VA bias. He further notes that even among the compliers, there was

the Chetty et al. evidence suggested that that wasn't true; some teachers are becoming more effective over time, some teachers are becoming less effective over time. And so he changed the bias question to ask whether models from this year predict next year’s impacts as well we would expect if the teachers are becoming more or less effective over time, and there was no bias beyond that. So in other words he redefined what previous research would have called bias as non-bias, and once you redefine that, he found what he calls evidence of non-bias.” (Mar 6 PM, 0h42m).
evidence of bias in English Language Arts (Mar 6 PM, 0h51m). Ultimately he feels that although MET is an important study, non-compliance precludes the researchers from being able to generalize their findings.

The next section of Rothstein’s remarks focuses on volatility in Value-Added scores, and is based on an article he co-authored with Linda Darling-Hammond and others (Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012).222 Rothstein reiterates that the reliability issues with VA scores undercut the face validity of these measures and opens up the teacher evaluation process to legal challenges from teachers. Based mostly on his own work (Rothstein J., 2010), he states that teacher effects as measured by VA are subject to dramatic fade out after the first year (Mar 6 PM, 0h58m-1h02m). He is also asked to evaluate Kane and Chetty’s conversion of test score impacts into “months of learning”; he contends that this measure is misleading because “nobody has ever actually done the study of ‘what is the impact of keeping the kids in school for two more months or two fewer months’.” (Mar 6 PM, 1h09m).

The capstone of Rothstein’s testimony came at the very end, and did not rely on simulation, but on an extant counterfactual to the challenged tenure statutes: charter schools. Rothstein testified that if the challenged statutes were harming student achievement, then charter schools – whose teachers generally do not benefit from those statutes – would out-perform traditional public schools. However:

“…they don't on average perform better than the traditional public schools, and that suggests that these regulations are not the cause of less than desired performance in the traditional public schools.” (Mar 6 PM, 1h58m-59m)

222 This paper’s contentions about the volatility of VA scores has already been discussed in Chapter 2.
The plaintiffs’ cross-examination built on their frustration with Rothstein’s simulation methods.\(^{223}\) The plaintiffs attempt to undercut Rothstein’s testimony by showing that despite his claims about teacher labor force policies, he lacks empirical evidence about California’s policies. They also get Rothstein to testify to the value of VA models for research purposes, including his own. Nonetheless, they were unable to unseat the charter school evidence, leading to them to call another witness as a rebuttal.

Expert Testimony: Eric Hanushek

Eric Hanushek was called by the plaintiffs in a rebuttal. Specifically, he was called to rebut expert witness testimony offered by the defendants – specifically Jesse Rothstein.\(^{224}\) His testimony shows the links between many concepts and events presented in this and other chapters. Indeed, the defense refers to Hanushek as “the father of Value-Added.” Hanushek was an expert witness in the 1973 \textit{Serrano v. Priest} case, which was the precursor to \textit{Vergara} in relating school funding to systemic inequality. In that case, he testified for the defense (California Public Schools) that teacher effectiveness mattered far more than school funding in

\(^{223}\) The plaintiffs’ cross-examination shows the difference that highly-paid counsel makes. Every point in their cross-examination was researched to correspond with Rothstein’s pre-trial deposition. By contrast, defense cross-examinations seemed far less planned. Indeed, the strategy of undercutting Rothstein’s simulation study could have been used against Chetty, who relied on data from New York, and Kane who relied on an RCT with high non-compliance.

\(^{224}\) The defense objects repeatedly to the plaintiff’s use of Hanushek as a rebuttal witness, arguing the plaintiffs were simply trying to get the last word, rather than to make an authentic rebuttal. The defense attempted to undercut the validity of value-added assessments as a teacher evaluation metric, just as the plaintiffs attempted to establish its validity in order to make the case of harm done to students by California teacher tenure policy. The judge ultimately admitted Hanushek as a witness, asking the plaintiffs to focus on specific testimony of defense witnesses (Mar 24 AM, 0h36m-39m). For the purposes of my argument, the rebuttal shows that \textit{Vergara} was as much about Value-Added methodology as about California teacher tenure.
determining student achievement (Mar 24 AM, 0h15m-16m).\textsuperscript{225} He served as an expert witness in twenty school funding equity cases since the 1970s; his amicus brief in Horne v. Flores, an Arizona funding equity case from 2009, was cited by Justice Samuel Alito in the Supreme Court’s majority opinion.

The theme of Hanushek’s remarks on VA was that the defense’s case made value-added out to be flawed for use in high stakes decisions, but that this misstates the purpose of VA altogether. Hanushek summarily refutes the various critiques of VA offered by defense expert witnesses (volatility of measures, bias, perverse incentives, etc.), saying that these criticisms are “irrelevant to some uses of value-added, and misleading with regards to others.” (Mar 24 AM, 0h26m). For example, he argues that economists use VA to gauge the difference between average teachers and ineffective teachers in the aggregate, not to evaluate individual teachers, and that this “main use” is unaffected by bias in VA measures (Mar 24 AM, 0h28m). In essence, Hanushek attempts to refute criticism of the policy uses to which VA is being put by reasserting the research purposes for which he created it. In principle, this is an acceptable position. However given the focus of the case and the hypothetical VA-based policy which Kane and Chetty rely upon to establish specific harm to students, this is misleading testimony.

Hanushek’s first substantive testimony regards his claim that firing the bottom five percent of value-added teachers would bring US educational attainment up to the level of Finland. He testifies that his five percent proposal implies using multiple measures for high stakes decisions, and that he would not support a system based only on VA (Mar 26 AM, 0h40m). However, his published work on this policy proposal is based solely on Value-Added measures of teacher

\textsuperscript{225} More recently, Hanushek wrote a book on the subject of funding and achievement (Hanushek & Lindseth, 2009).
effectiveness, and actually advocates firing between five and ten percent of teachers each year (Hanushek E. A., 2009).

The next important point he makes is a response to earlier testimony by Jesse Rothstein that firing teachers would not generate the achievement gains promised by Hanushek and others (Staiger & Rockoff, 2010). This is based on Rothstein’s own simulation study (2012), which was later published in the *American Economic Review*. Hanushek critiques Rothstein’s study by arguing that:

“...economists sometimes use very simplified models to study different outcomes, and that’s what this is, and in fact a constructed model to show a specific thing in this case…perhaps the best analogy is that if I sat here and put a rabbit into my hat and then turned around and reached into my hat and pulled out a rabbit, you shouldn’t be surprised. He has constructed a model that tries to relate characteristics of teacher contracts to a stylized version of the teacher labor market and their implications for student outcomes, but it takes a strong a specialized view which has no bearing on the reality that we’ve seen…” (Mar 24 AM, 0h52m-53m)

Ironically, much early work on Value-Added relied on simulation data, and the very same critique of simplified modeling and circular logic can and has been leveled at past and present VA research. Hanushek further argues that Rothstein is wrong about the teacher labor market, stating that there is a “reserve army of people looking to enter teaching.” He criticizes Rothstein’s reliance on the Tennessee STAR experiment to infer class size effects on achievement, since the effect is only found in Kindergarten and first grade and other studies of class-size reduction show no effects on achievement (Mar 24 AM, 0h57m). But the STAR experiment was a large random assignment study (the largest educational experiment ever prior to the MET project), which according to many analysts found significant effects across grade levels. For Hanushek to diminish its findings would be akin to dismissing Kane’s study on Value-Added. On the issue of teacher-pupil ratio, Hanushek’s research takes Rothstein’s role,
picking apart the evidence by meticulously mapping out the educational context (Hanushek E. A., 1999). The plaintiff’s end their examination of Hanushek on this point.\footnote{On cross-examination, Hanushek states that he views the results of the STAR experiment to be unreliable, and that although he objects to Jesse Rothstein’s use of its findings in his study, he does not object to similar use of STAR data by Raj Chetty et al. in \textit{Long Term Impacts} (Mar 24 AM, 1h31m). Curiously, the defense attorney does not pursue this point any further – given that it could characterize Hanushek as politically driven in his interpretation of research.}

The teachers’ union’s cross-examination of Hanushek first focuses on VA models. The defense establishes that Hanushek sees potential issues with the use of VA models, particularly the possibility that reliance on student achievement gains may lead to ‘teaching to the test’ or cheating (Mar 24 AM, 1h25m). They then establish that Hanushek views other factors – schools, neighborhoods, peers, families – as pertinent to student achievement (Mar 24 AM, 1h29m). In terms of the teacher labor force question, the defense unsuccessfully attempts to establish that Hanushek’s five percent policy would increase the risk of working as a teacher.

The next piece of the cross-examination focuses on a series of studies by the Center for Research on Educational Outcomes (CREDO) which compares charter school and public school performance. In direct examination, Hanushek referred to a CREDO report on Los Angeles charter schools to argue that charter schools outperform traditional public schools (Center for Research on Educational Outcomes, 2014a). This was in response to Jesse Rothstein’s testimony that changing the teacher tenure policy could harm student performance. Since charter schools have much less robust tenure rules, Hanushek uses the LA report to claim that this doesn’t necessarily hold true. The defense challenges Hanushek’s credibility by citing contrary evidence from a statewide CREDO study of California charter schools (Center for Research on Educational Outcomes, 2014b). Hanushek says this report did not inform his opinion on the
subject because he was not familiar with it because it was only recently published (Mar 24 AM 2h05m). This is a dubious claim; upon further questioning, we find out that both studies were primarily authored by Hanushek’s wife, CREDO director Margaret Raymond. Although not mentioned in the case, the timing of the report was clearly not an issue either; the California CREDO report was published only one day after the Los Angeles report. The defense attempts to establish the same credibility problem by referring to other CREDO studies with similar, but is unable to successfully connect those studies to Hanushek’s testimony. The defense’s last attempt at this strategy cites on Hanushek’s long one-sided history as an expert witness in school funding equity cases, but the court does not allow this line of questioning (Mar 24 PM, 0h39m). After a brief re-direct examination, Hanushek leaves the stand and the defense rests.

Expert witness testimony in Vergara proved that the trial was as much about Value-Added methods as it was about teacher tenure. In attempting to establish specific harm done to students by California’s tenure policies, the plaintiffs posited value-added as a counterfactual mode of assessment. Their witnesses, Raj Chetty in particular, converted teachers’ hypothesized impacts on student test scores into months of learning lost and million-dollar income shortfalls. They further claimed that using value-added in retention, promotion and dismissal policies would mitigate the harm done to students by ‘grossly ineffective teachers.’ In doing so, they had to

---

227 Hanushek founded CREDO in 1998 at the University of Rochester before relocating it to Stanford in 2000.

228 Seeking to get the last word, the plaintiffs re-examine Hanushek. They confirm that his testimony regarding charter schools does not affect his opinions regarding the effects of teachers. Like Thomas Kane, Hanushek translates student test score gains into days of learning (Mar 24 PM, 0h48m-49m). The re-direct examination ends after referring to a number of charts from the CREDO Los Angeles charter study which focus on minority student achievement (Mar 24 PM, 0h50m-55m). The defense attempts to re-examine Jesse Rothstein in rebuttal, but his testimony is not admitted by the court after successful objections by the plaintiffs (Mar 24 PM, 2h19m-2h34m).
validate value-added before the court. This validation relied on two major trends in social policy science. Chetty’s *Long Term Impacts* study leveraged so-called Big Data; and Kane’s *Measures of Effective Teaching* project featured the largest random assignment study ever conducted in education research. As I have shown, both Chetty and Kane presented their research unburdened by important caveats that have been raised by critics. As an expert for the defense, Jesse Rothstein gave voice to these criticisms, troubling the foundation of the plaintiff’s case. Rothstein contends that using value-added methods beyond their intended research purpose for high-stakes decisions has the potential to do harm to school systems. To challenge the testimonies of Kane and Chetty, he argues that such measures are too biased and too volatile to be used as an alternative to existing teacher evaluation and tenure systems.\(^{229}\) He also demonstrates that existing tenure rules could not be the cause of poor student achievement, since charter schools are not subject to those rules and achieve similar results. To rebut his testimony, the plaintiffs relied on Eric Hanushek, the “father of Value-Added” and a veteran of politically-charged education research, to buttress their argument. After another day of closing arguments and two months deliberation, the judge ruled.

Outcome and Press Coverage of Vergara

Oral arguments in *Vergara* ended on March 27\(^{\text{th}}\), 2014 and the case officially concluded on April 10\(^{\text{th}}\). Judge Treu handed down a 16-page ruling on June 10\(^{\text{th}}\). He ruled in favor of the plaintiffs that the “challenged statutes” (seniority-based reductions in force, the two-year probationary

\(^{229}\) Rothstein’s testimony was weakened by the fact that some of his research relies on simulations. In the middle of direct examination by the defense, Mr. McRae argued before the judge: “Your Honor, the difference with all of our experts is that their studies were actually based on DATA [my emphasis] where they went in and did countless numbers of examinations with actual students and actual test scores and wasn't again just plucked as 'in vacuo' an abstraction 'if you do this, will it have that effect?'” (Mar 6 PM, 1h14m-15m)
period for tenure, and various due process protections) caused students to be exposed to ‘grossly ineffective teachers’ and thus violated their constitutional rights to a good education. Though much of the decision relies on testimony on teacher dismissal processes, its definition of grossly ineffective teachers relies on Value-Added studies conducted by Chetty and Kane. Treu’s first and most dramatic point refers to these studies:

“Evidence has been elicited in this trial of the specific effect of grossly ineffective teachers on students. The evidence is compelling. Indeed, it shocks the conscience. Based on a massive study, Dr. Chetty testified that a single year in a classroom with a grossly ineffective teacher costs students $1.4 million in lifetime earnings per classroom. Based on a 4 year study, Dr. Kane testified that students in LAUSD who are taught by a teacher in the bottom 5% of competence lose 9.54 months of learning in a single year compared to students with average teachers.” (Vergara v. State of California, 2014, p. 7)

Despite the many technical nuances of the expert testimony, the ruling is curt on the question of identifying ineffective teachers. Treu assumes that grossly ineffective teachers exist and ascertains a number of them. The sparring between Jesse Rothstein and Eric Hanushek over the performance of charter schools – though important to both sides during trial – was not mentioned in the ruling. Indeed, despite his lengthy stay on the witness stand, Rothstein’s testimony is not referred to once, even negatively. The judge’s very short opinion – only 4,000 words long – drew criticism from the teachers’ unions who intervened in the case. They charged

---

230 The frequency of grossly ineffective teachers in Los Angeles is based on the testimony David Berliner, another defense expert. In cross-examination, Berliner estimated that somewhere between 1 and 3 percent of active teachers are grossly ineffective. This figure was extrapolated in Treu’s ruling to a range of 2,750 – 8,250 teachers (Vergara v. State of California, 2014, p. 7). Berliner later argued that his testimony on the proportion was mischaracterized in the ruling (Murphy, 2014).
that it lacked any thorough or reasoned consideration of either the statutes in question or the larger context in which they function (Weingarten, 2014; Weingarten & Pechthalt, 2014).

The Vergara ruling received much press attention, which I now turn to analyze. I collected 53 documents in all: 17 stories came from major news outlets within two days of the ruling, eight editorials and opinion pieces from those same sources, 15 stories from high-visibility online sources (like The New York Times and Washington Post blogs, and online magazines Slate and The Atlantic), and 13 additional pieces from up to two weeks after the ruling and/or from less visible news outlets. I consider what aspects of the ruling and the reaction received the most attention; I then conclude by assessing how value-added was translated from expert testimony to ruling to press coverage.

News media love good headlines. In his ruling, Judge Treu used two key phrases (both in the quoted paragraph above): ‘grossly ineffective teachers’ and, referring to the impacts of said teachers, ‘shocks the conscience.’ These were by far the most oft-cited phrases in the news coverage. 20 of 25 major news stories and editorials included the first phrase; 14 of 25 included the latter. Of all 53 pieces, 30 used the former and 22 the latter. Further, a number of stories repeatedly used the grossly ineffective teacher phrase. Across articles from major news outlets, some version of grossly ineffective was used 30 times; in all 53 pieces, the phrase appears 60 times.

---

231 The ruling was very short when compared to other education law cases. The 1971 Serrano v. Priest opinion was 15,000 words long; the opinion in the 1976 iteration of that case was 28,000 words. The 2010 Doe v. Deasy opinion was approximately 10,000 words.

232 I coded and analyzed these documents using Atlas-ti qualitative data analysis software.
These ‘sound-bites’ from the ruling are used rather uncritically. The phrase ‘grossly ineffective teachers’ was used by the plaintiffs throughout the trial and is directly related to their reliance on value-added – grossly ineffective means the bottom five percent of the VA distribution. Treu’s conscience-shocking evidence also relies on VA measures; it refers to the suggested difference between being assigned a bottom 5th-percentile teacher as opposed to an “average” teacher. Eight of the 25 major outlet pieces (and 10 stories overall) also referred to Treu’s potentially dubious estimate of the number of grossly ineffective teachers. Yet only one major new story (and six overall) make any mention of value-added methods. Similarly, few stories mention Raj Chetty or Thomas Kane – the researchers who produced this evidence – despite their names being cited in the judge’s opinion. Expert witnesses for the defense – Jesse Rothstein and David Berliner – were each mentioned in only three stories out of 53.

As would be expected, the challenged employment statutes in Vergara received more attention than value-added; but this was largely driven by a few data points. LA Schools superintendent John Deasy was the most often referenced witness, mentioned in seven of the major outlet stories and 12 pieces overall. This is likely because the case was about California tenure rules and because Deasy testified about the potential cost of teacher dismissal. The figure cited in the ruling – $50,000 to $450,000 per teacher – was referenced in news coverage more often than Deasy himself (nine major outlet stories). California’s two-year probationary period was mentioned in 14 out of all 53 pieces. Nonetheless, none of the actual substance of the trial received as much attention as did the quotable phrases discussed above.

The two parties to Vergara – Students Matter and the two California teachers’ unions – received somewhat similar amounts attention. Nearly all the major news stories mentioned Students Matter (17 of 25) and at least one involved union (19 of 25). However, only 9 of 25 news stories
gave background on Students Matter, which is an important omission given that the organization financed the case. Given that the ruling was in favor of the plaintiffs, 12 of 17 major news stories (not including editorials) covered the statements and press releases from the teachers’ unions.

Judge Treu’s terse ruling may be partly responsible for the lack of depth in press coverage following *Vergara*. There was simply not much information to report in the 16-page opinion. But given the potential impact of the ruling, and the fact that both plaintiffs and defendants maintained extensive online presence throughout the 3-month long trial – providing links to both admitted evidence and witness testimony – there was reason and opportunity to expand coverage. Very few stories, even in California, gave readers substantial background, specifically on the concept of ‘grossly ineffective teachers.’ Without context, the value-added definition of effective and (grossly) ineffective teaching is reified by both the ruling and the subsequent press coverage.

This chapter has shown how value-added, a methodologically-contested technique, is deployed tactically and simplistically in a public forum like a trial. Despite defense experts introducing the complexity and contingency of this technique, the judge was persuaded by exaggerated and debatable data points. Thus value-added data – complex and fraught as it was – was simply transformed into ‘grossly ineffective teachers’ whose negative impacts ‘shocked the conscience.’ When *Vergara* was covered in the press, this evidence was turned into ‘sound byte’ coverage with little or no context. In the conclusion, I will consider the implications of this translation for theorizing the relationship between technology and policy in education.
Chapter 7: Understanding Value-Added: Reason, Rationality and Politics in the Social Sciences

The preceding chapters have made two broad arguments. The first was to show that the assumptions of value-added assessments (VA) are at odds with many of the practical realities of the schools and classrooms these techniques are intended to evaluate. The second was to argue that this disconnect threatens the intended social functions of education. To substantiate the divergence between VA in theory and schools in practice, I demonstrated three empirical problems with VA measurements using data from New York City.

Second, by examining the history of innovations in US educational assessment, I argued that the politics of school reform are impacted by technical ‘upgrades,’ particularly as they coincide with claims of educational decline and consequent national peril. To understand VA in this context, I traced the history of its technical evolution and related this to concurrent political developments in the educational sphere. Making use of expert testimony in and press coverage of Vergara v. California, I showed how VA research on teacher effectiveness was simplified to make VA methods appear objective and neutral and then stylized to fit a political agenda of educational privatization and top-down management.

My intention in this project has been to use value-added assessment as a case study in the relationship between technology – defined as both research methodology and the tools employed in research – and social policy. To conclude, I return to social theory to consider different perspectives on the connections between social science, technology and politics, specifically as they relate to education. More simply, I ask under what conditions and by what mechanisms does education come under the influence of often unsound science and technology?
Two broad positions must be explored. The first argues that science and technology, regarded neutral and objective, are corrupted and deliberately deployed by actors in the political sphere. The second contends that ideology and power are inherent in science and technology. This latter position implies that technology need not be corrupted or politicized, as its very application contains \textit{a priori} ideological assumptions. Let us begin by consider the theoretical underpinnings and specific examples of each position.

Value-Neutral Science

Understanding the value-neutral view of science requires going back to Kant’s \textit{Critique of Pure Reason}, in which he argues that reason corresponds to a class of knowledge which is independent of judgment or experience.\textsuperscript{233} In outlining what part reason plays in our knowledge of reality, Kant describes mathematics and the laws of natural science (by which he primarily meant physics) as capable of producing \textit{synthetic a priori} knowledge – \textit{synthetic} in the sense that they combine two or more concepts and \textit{a priori} in the sense that each constituent concept is defined axiomatically. The statement “$8+3=11$” is such a piece of knowledge, as ‘8’ and ‘3’ are always what they are, and their sum is a synthesis of the two.\textsuperscript{234} But Kant, deeply influenced by the skeptical tradition of Hume, is careful to say that our knowledge of the natural world is necessarily limited. Since experience is mediated by our faculties of sensibility and

\textsuperscript{233} One can indeed trace the value-neutrality question much further back. Walsh (2013), quoting Samuel Coleridge, contends that the fact/value, is/ought divide originates with Plato and Aristotle. Though he traces the general spirit of philosophy back to the Greeks, Husserl (1954/1970) credits Descartes with the earliest claim to this legacy.

\textsuperscript{234} Kant’s view of science is nuanced and contained in many works over his career. What is important in the context of this discussion is that Kant’s definition of science as distinct from ‘systematic arts’ is very restrictive and almost entirely bound up with the use of mathematics. Thus his famous line in \textit{Metaphysical Foundations} “in any special doctrine of nature there can be only as much proper science as there is mathematics therein.”
understanding, which organizes sensory experience according to intuitions and principles (e.g., cause and effect), Kant argues we are unable to know the actual contents of the universe axiomatically – the things themselves. Thus, while Kant is appropriately careful to delimit our knowledge of the external world, the idea that knowledge can be generated independent of experience (through reason alone) opens up space for a value-free science.

Rationalist and positivist philosophy since Kant, itself influenced by scientific and technological advance, argued that the world in itself is indeed objectively knowable. Combining Hume’s empirical principle that all we can know is the sense data of experience with the rationalist view that reason is primary in acquisition of knowledge, logical positivism – associated with the early 20th century Vienna Circle – asserts that knowledge of reality is possible through the construction of verifiable logical statements based on empirical observation. This philosophical perspective stands in opposition to any metaphysics and implicitly to the mediated experience of the world proposed by Kant (and by Hegel in *The Phenomenology of Spirit*). This rejection of metaphysics furthers the possibility of a value-free science by asserting that empirical observation reliably describes the world as it is.

But positivism as a worldview was formulated earlier by the ‘father of sociology,’ Auguste Comte. Relatively obscure other than for his coining of the terms ‘sociology’ and ‘positivism,’ his influence is largely ignored or disavowed by social scientists whose approach was anticipated by Comte’s program for positive philosophy (Lenzer, 1998).

Comte was deeply concerned with the revolutionary political climate which characterized his times – which he called moral and political anarchy. In response to this ‘crisis,’ he devoted his life to using the tools of science to facilitate an orderly society. Contrary to the liberal
philosophers of the time, Comte located the source of social unrest in an excess of individual freedom (be it of a king or the revolutionary masses) rather than the tyranny of government. Specifically, he felt that the critical spirit of political philosophy needed to be replaced with a positive spirit. Comte thus sought to resolve the conflict between the individual and society in favor of society, and specifically in favor of non-arbitrary power grounded in positive political science. In the service of this project, he sought to found an objective ‘social physics’ which would allow the “preponderance of observation over imagination” in political affairs (Comte, Plan for the Scientific Operations Necessary for Reorganizing Society, 1822/1998, p. 36). Comte further sought to directly attach his efforts to those in power – offering his services to the Russian Czar Nicholas I and to the Ottoman Emperor (Pickering, 2009, pp. 77-80).

While this context sounds like the opposite of value-free, Comte’s positive social science was to be free of all “admiration or reprobation of phenomena;” while he refers to the critical spirit of liberal philosophy as “directly contrary to that which ought to reign in scientific politics” (Comte, Plan for the Scientific Operations Necessary for Reorganizing Society, 1822/1998, p. 54). Comte details his philosophy of science in the six-volume Cours de Philosophie Positive. Reflecting his value-free ethos, and anticipating later developments in social science methodology, he argues that social science must emulate the techniques of the natural sciences and places mathematics at the foundation of all other sciences:

“In the present state of our knowledge we must regard mathematics less as a constituent part of natural philosophy than as having been, since the time of Descartes and Newton,

\[\text{235}\]

Comte’s early formulation of his positive political science suggests collaboration with the existing ruling classes:

“The governing classes, clearly perceiving [through positive knowledge of social science] the end that they are called on to realize, can reach it directly…They will by anticipation combine measures for overcoming opposition with others calculated to facilitate the acceptance of the new order of things by their opponents.” (Comte, 1822/1998, p. 46)
the true basis of the whole of natural philosophy…as being the most powerful instrument
the mind can employ in the investigation of the laws of natural phenomena. Geometry
and mechanics must, on the contrary, be regarded as true natural sciences…But these two
physical sciences…will be more and more, employed rather as method than as doctrine.”
(Comte, View of the Hierarchy of the Positive Sciences, 1853/1998, p. 100)

Comte’s subsequent move toward a secular religion based on his social and political theory was
rejected by most of his contemporaries – and helps explain his obscurity in modern discourse on
social science methodology. Nevertheless, Comte’s spirit of detached, objective inquiry into
social phenomena was extended by canonical sociological thinkers, perhaps the most significant
of whom is Emile Durkheim. Durkheim’s (1895) *The Rules of Sociological Method*, in which he
tries to carve out a distinct space for sociological inquiry, rests on Comte’s value-free, anti-
metaphysical, and rationalistic principle. His position on the role of social scientists in
political affairs is that they should play, if anything, an advisory role (Lukes, 1982). Durkheim
does however, move from Comte in his preferred techniques of inquiry and assumptions required
for the study of ‘social facts.’

Max Weber (1864-1920) was thoughtful about the role of social scientist in society and made an
attempt to distinguish the objects of social scientific study from the value they imply. In his
influential 1904 essay *Objectivity in Social Science and Social Policy*, Weber begins with the
observation that “an empirical science cannot tell anyone what they should do” (Weber,

---

236 Recognizing some limitations, Comte did note that social physics – only later renamed
sociology – was not yet fit for mathematical analysis, if only because mathematical education
was still inadequate to produce the models needed for these “difficult researches.” His ultimate
vision is that mathematical reasoning will absorb and eliminate “the only part of the old
philosophy that could even appear to offer any real utility, the logical part.” (Comte, 1853/1998,
p. 177) Thus we see that Comte’s positivism sought to create a value-free social science through
the application of natural scientific and ultimately mathematical methods, placing him very much
in the tradition which informs much of contemporary social science.

237 Indeed, Durkheim cites Comte’s *Cours de philosophie positive* as “the only original and
important study which we possess on the subject [of sociological method]” (Durkheim,
1904/1949, p. 49). In an apparent reference to Comte, he declares that “only positive religions – or more precisely expressed, dogmatic sects – are able to confer on the content of cultural values the status of unconditionally valid ethical imperatives (p. 57). From these basic principles, Weber suggests that what we now call ‘policy science’ is impossible because problems of social policy cannot be resolved based on facts alone, and indeed because policy decisions must be worked out based on debating cultural values (p. 60).

Weber thus acknowledges a fact/value distinction which constantly presents itself to the social scientist. The objects of social science are inherently meaningful for Weber, thus precluding the possibility of ‘presuppositionless’ investigation (p. 76). Given this, he lays out two primary responsibilities for the social scientist. The first is that one must always make clear one’s value judgments in two ways – specifically the standpoint from which one’s analysis departs and moments in writing when discussion shifts from scientific analysis of fact to evaluation of policy (pp. 59-60). The second responsibility is that one must attempt to remain value-neutral in analysis as much as possible. In practice, what this entails is paying attention to evidence which contradicts the dominant formulations or analytic constructs that one uses in analysis.

This latter imperative is connected to Weber’s main methodological construct: the ideal type. Weber demands that social scientists pay attention to empirical events which do not agree with the ideal types which drive their formulations. He further points out how tempting it can be to “do violence to reality in order to prove the real validity of the construct” (p. 103) and how scientists can often confuse their analytic ideal types with their own subjective ideals (p. 98). But he nonetheless holds out the possibility that social science can use value-laden standpoints as analytic tools while refraining from using them as evaluative principles. Thus, although he holds
a more pessimistic view, Weber still maintains that social science can be distinct from any particular value set.\textsuperscript{238}

Holding aside for the moment the criticisms of positivism beginning in the early 20\textsuperscript{th} century (these will be discussed below, as they constitute the basis of the other broad position), the idea of value-free, politically detached science continued under the auspices of ‘post-positivism.’\textsuperscript{239}

As a response to various critiques, post-positivism concedes that empirical observation and analysis is necessarily “theory-laden,” and that we cannot attain certain truth through scientific inquiry, only (and at best) conjectural knowledge (Trochim, 2006). Regardless of this epistemological position, post-positivism methodologically produces the same purportedly ‘objective’ knowledge claims as its positivist forbears. Further, the philosophical move to contingent knowledge did not prevent the growing emphasis on mathematical and analytic methods which increasingly shaped social science in the mold of the natural sciences. Phillips and Burbules (2000) note that the philosophical criticisms that turned positivism into post-positivism are largely ignored in texts on research methods, and thus implicitly turn observation and empirical data into unmediated (and value-free) truth.\textsuperscript{240} Introducing a volume of essays on positivism and alternative epistemologies across the social sciences, Steinmetz (2005) contends

\textsuperscript{238} This view did not prevent Weber from holding social scientists accountable for making political choices. But he insisted that these choices should be made specifically in their roles as citizens, since social scientists “are the least fitted to presume to save the individual the difficulty of making a choice, and they should therefore not create the impression that they can do so.” (Weber, 1949, p. 19)

\textsuperscript{239} Post-positivism is elsewhere called anti-positivism and non-positivism; my use of the former is arbitrary.

\textsuperscript{240} One popular research methods textbook which features such an elision was co-authored by Eric Hanushek, whose highly politicized research career has been discussed throughout the preceding chapters (Hanushek & Jackson, 1977). Hanushek’s introduction singles out Economics as more rigorously deductive than other social sciences, but even contemporary methods texts in sociology present the relationship between theory and empirical data as untroubled by social forces (Healey, 2012).
that despite varying trajectories depending on the discipline, social science methods are still in the midst of a “positivist haunting” (p. 3).

This omission from texts on methodology partially reflects a pragmatic orientation on the part of researchers, who ignore questions of objectivity and ultimate truth for the sake of ‘getting down to business’ (for examples in education research, see Miles & Huberman, 1984; Howe, 1988). A related explanation for is that as research became more policy-oriented (and government funded) over the course of the 20th century, the assertion that the ‘normal’ functioning of society constitutes empirical fact was elevated from theoretical claim to required assumption. Mirowski (2005) traces the relationship between logical positivism and operations research (OR) in the United States and argues that the immense funding and relative autonomy afforded to researchers working on defense contracts came at the cost of questioning the uses to which their work was put. Indeed, the trajectory of educational research methods traced in the preceding chapters exemplifies the ascendancy of the postpositivist, analytic paradigm and the extent to which claims of objectivity are taken for granted.

The dominant assumption of a value-free scientific enterprise is also reinforced by the mainstream sociology of science, which largely eschews questions of power or interest. Robert Merton’s program for the sociology of science focused on empirical markers of scientific activity – publications and biographies – and on reward systems within the sciences, rather than

---

241 Weber (1904/1949) would contend that this sort of ignoring of value judgment cannot lead to value-neutrality, but only to unacknowledged value-judgments which nonetheless make their way into the evaluation of the ‘facts’ (for example see pp. 82,94).

242 Both Comte and Durkheim express versions of the ‘normative is natural’ claim. Comte argues that society as it currently exists is a manifestation of fundamental laws that govern the progress of civilization (Comte, 1822/1998, pp. 42-44). Durkheim’s fundamental distinction between the ‘normal’ and the ‘pathological’ is a more measured version of the same concept.
examining at how scientific findings are deployed outside the profession, or on how outside forces determine what is studied (Merton, 1973). Related tendencies in the history and philosophy of science focus on its internal dynamics – that is, how scientific knowledge accumulates and changes over time. Karl Popper’s (1959/1968) *The Logic of Scientific Discovery* and Thomas Kuhn’s (1962) *The Structure of Scientific Revolutions* are among the best known works in this tradition. But both the sociology and history of science largely take for granted the isolation of scientific research from the corrupting influence of politics. Whatever the causes, and despite important critical upsurges – variously named ‘Science Wars’ and ‘Paradigm Wars’ –, versions of positivistic practice claiming value-free status are generally dominant across the social sciences (Steinmetz, 2005a).

Science as Value-Laden

Positivism and its epistemological descendants all rest on the assumption that scientific inquiry can describe the world as it is without consequential mediation by human values. But this position gave rise to many criticisms which claim that science immanently contains its own value system. We even see a germ of this critique in Weber, though he addresses this primarily to

---

243 In a later essay which appeared in the same volume, Merton does attempt to outline the relationship between applied research and policy (Merton, 1973a). Among the insights offered in that paper, Merton does attempt to understand how policy research is utilized, and includes the possibility of co-optation of research by politicians (p. 75).

244 Kuhn’s description of how knowledge accumulates does differ from positivistic accounts of scientific progress. Once a ‘paradigm shift’ occurs, Kuhn argues that the language used to describe objects of study changes so drastically that the facts discovered under one paradigm are incommensurable with those of another. Hence Kuhn’s critique is placed within the anti-positivist tradition. But demonstrating its isolation from politics altogether, Kuhn’s account does not even consider the role of academic politics in thwarting important theoretical developments. Looking at String Theory in physics, Lee Smolin (2008) argues that the failure of much work in physical science is due to boundary maintenance and careerist tendencies among university physicists.
clinical science (medicine) and technology (engineering), arguing that these applied sciences assume evaluative standpoints in advance and proceed with specific goals in mind – patient health in the former case and productive efficiency in the latter (Weber, Objectivity in Social Science and Social Policy, 1904/1949, pp. 85-86).

Significant critiques of positivist social science arose from phenomenology, the Frankfurt School and later from French social theory. Each in its own way seeks to reassert the subjective dimension of scientific inquiry and the technology it produces.

Edmund Husserl’s Vienna Lecture\(^{245}\), delivered in 1935 characterizes his attempt to account for the failure of European Sciences (and more broadly of European humanity), which he believed had fallen victim to a positivistic spirit which had “decapitated philosophy” (1954/1970, p. 9) and left even the positive sciences (e.g., physics and mathematics) in a crisis of meaning despite their practical accomplishments. This lecture also outlines his phenomenological approach to knowing and understanding the world.

Husserl accepts the successes of the physical and biological sciences and attributes these successes to the application of theoretical reason to empirical reality and to the search for universal laws, “the result of [which]…was a true revolution in the technical control of nature” (Husserl, Philosophy and the Crisis of European Humanity, 1954/1970, p. 271). But he argues that humanistic disciplines cannot seek to uncover such natural laws because humanist objects –

\(^{245}\) I take Husserl as representative of the phenomenological critique of positivism and value-free social science. The Vienna Lecture was the precursor to Husserl’s final book-length attempt to outline his phenomenology, *The Crisis of European Sciences and Transcendental Phenomenology*. I refer to this (both the Vienna Lecture and *The Crisis*) among his works because it speaks directly to the theme of this chapter. Other phenomenological critiques are certainly worth considering (see for example, Merleau-Ponty, 1964/1968).
Freud’s psyche or Weber’s social values – are bound up with individual subjective experience. Abstraction from nature (rather than from spirit) thus cannot lead the humanist to an exact science.

To explain what he calls the crisis, Husserl first traces the history of the European Sciences back to Ancient Greece. The worldview which originates with the Greeks is the pursuit of infinite, unconditional and durable ideas – what Husserl calls the theoretical attitude (p. 280). This contrasts with other forms of human pursuit which although like science, were concerned with successful production, did not seek infinite truths, only more immediate ends. He calls this the natural or practical attitude. But it is the further transition from the theoretical attitude of philosophy (for example, Plato’s disinterested pursuit of the forms) to that of science (which applies ideal norms to empirical matters) which for Husserl constitutes the decisive shift, not only among the scientific community, but also in the general population through scientific education (pp. 285-287).

The stumbling block which turns this historical progression into a crisis lies in what Husserl calls naïve rationalism, ‘naturalism,’ or ‘objectivism.’ While acknowledging that scientific pursuit requires systematic focus on specific objects, Husserl contends that this periodic focus must be

---

246 This is the essential implication of Husserl’s concept of the lifeworld. Rather than there being some objective world that we all experience in exactly the same way, Husserl argues that what we regard as the external world occurs to us as phenomenon. This impression of the world can be socially founded, and it is precisely a malaise of this impression which Husserl exposes in the Vienna Lecture.

247 Husserl defines Europe not in a geographical sense, but in the spiritual sense of a shared experience of the world (lifeworld). This leads him to include the United States and other British and European colonies in the concept, as they share European institutions and societal goals (Husserl, Philosophy and the Crisis of European Humanity, 1954/1970, p. 273). He distinguishes Greek philosophy from earlier wisdom (Egyptian, Babylonian)
tempered by reflection.\textsuperscript{248} He argues that the discoveries of math, geometry, and physics were so successful (in producing tools and predictions) that without appropriate reflection mathematical/geometrical reason was applied to the whole of natural and psychic (psychological and moral) experience. Thus, “in…the constantly objectivist attitude, everything spiritual appeared as if it were [simply] spread over [the surface] of physical bodes” (Husserl, 1954/1970, p. 291).\textsuperscript{249}

Further, the one-sided pursuit of natural and spiritual laws leads the scientist to forget the subjective experience which grounds his pursuit of universal laws and instead to leads him to idealize his experience of the world as ‘apodictic’ or objectively self-evident (Husserl, 1954/1970, pp. 294-297). ‘Objectivism’ is then the naïve belief that our individual experience of the world, however rigorously investigated, is unconditional and generally applicable. Husserl proposes transcendental phenomenology – a systematic inquiry into subjective experience – as the necessary groundwork for all natural and humanistic sciences to overcome the objectivism that characterizes the European sciences.

The pheonomenological critique offered by Husserl is fundamental; from it we understand the basic problem of naturalizing (making objective) subjective experience – it obscures value-laden choices made in the process of inquiry. But to probe the specific values which underwrite this objectification, we turn to the Frankfurt School critique of positive science. This critique is

\textsuperscript{248} “The true philosopher must always devote himself to mastering the true and full sense of philosophy, the totality of its horizons of infinity. No line of knowledge, no single truth may be absolutized and isolated…Only through this constant reflexivity is a philosophy universal knowledge” (Husserl, 1954/1970, p. 291).

\textsuperscript{249} In a separate essay, Husserl discusses in detail the assumptions required in the mathematization of nature. Therein he demonstrates that what we call assumptions are perceptions which we turn into logical (self-evident) propositions.
exemplified in Horkheimer and Adorno’s *Dialectic of Enlightenment* (1944/2002), but is grounded in the concept of ‘reification’ advanced by Georg Lukacs (1923/1971).

Like Husserl, Lukacs engages Kant’s problem of epistemology and critiques the subject-object split which characterized positive natural and social sciences in the early 20th century. A further similarity is that he shows how this split in the sciences affects the rest of social life. But Lukacs argues that this artificial split – especially in the social sciences – reifies human relations following the model of capitalist class relations. Specifically, he connects the emergence of objective science with that of the commodity form – arguing that the subjective stance of human social relations is fundamentally altered by the increasing dominance of the commodity form:

“Only in this context does the reification produced by commodity relations assume decisive importance both for the objective evolution of society and for the stance adopted by men towards it. Only then does the commodity become crucial for the subjugation of men’s consciousness to the forms in which this reification finds expression and for their attempts to comprehend the process or to rebel against its disastrous effects and liberate themselves from servitude to the ‘second nature’ so created.” (Lukacs, 1923/1971, p. 86)

Lukacs thus connects the critique of objective science with the Marxist critique of capitalism by implicating bourgeois values as the basis of inquiry. Frankfurt school theorists take his argument further as it relates to science and technology. Max Horkheimer (1932/1999) argues that science is limited by its existence as a mode of production under capitalism – with its products being social values and useful applications for industry. He regards modern science as superficial in its ahistorical focus on improving existing conditions and in its seemingly arbitrary – but in reality

---

250 Lukacs uses Kant as an exemplar of bourgeois thought which divides the world of thought and intelligible forms from the world of objects. He also connects the mathematization of natural and social phenomena (and the specialization which comes along with it) with the split effected by the commodity form (Lukacs, 1923/1971, pp. 110-121).

251 Reification as a concept is developed by Lukacs, but originally comes from Georg Simmel who first proposed it in his “Philosophy of Money.”
ideological – choice of research tasks (p. 211). Horkheimer’s prognosis for science is similar to Lukacs: since the crisis of science is “inseparable from the general crisis,” (p. 212) only a revolution in the broad social conditions within which science operates is capable of freeing science from its crisis.

This perspective on science is further explored in Horkheimer’s collaboration with Theodor Adorno on the concept of Enlightenment. Adorno and Horkheimer contend, similar to Husserl, that rationality is taken to an absurd extreme in the age of modern science. This overemphasis on rational (specifically mathematical) calculability reflects and reinforces a system of domination over nature, and promotes suspicion of alternative conceptions and data that stands beyond calculation. Like Lukacs they see the cultural dominance of science as the outcome of an historical development in which the rationalism of the Enlightenment had to displace the ideological dominance of religion. But they pessimistically regard the process as nearly complete; for them, enlightenment rationality has all but disposed of its epistemological others and become self-maintaining. This dominance has a corrosive effect on the critical power of enlightenment insofar as individuals model themselves on such standardized behavior and are unable to conceive alternatives to or criticisms of the existing order (Horkheimer & Adorno, 2002).

---

252 Horkheimer seems to address Husserl’s position directly when he says that: ‘Science in the pre-War years had a number of limitations. These were due, however, not to an exaggeration of its rational character [e.g., positivism] but to restrictions on it which were themselves conditioned by the increasing rigidification of the social situation.” (Horkheimer, 1932/1999, p. 209)

253 “The manifold affinities between existing things are supplanted by the single relationship between the subject who confers meaning and the meaningless object, between rational significance and its accidental bearer.” (Horkheimer & Adorno, 1944/2002, p. 7)

254 “Knowledge does not consist in mere perception, classification, and calculation but precisely in the determining negation of whatever is directly at hand…mathematical formalism…arrests thought at mere immediacy…The more completely the machinery of though subjugates existence, the more blindly it is satisfied with reproducing it.” (Horkheimer & Adorno, 1944/2002, p. 20)
1944/2002, p. 22). In their analysis, the productive values inherent in science are already largely taken for granted in society and thereby further insulated from critical scrutiny.  

While Horkheimer and Adorno take the insights of Lukacs as a point of departure, they go beyond positing science as an agent of capitalist domination and afford independent ideological power Enlightenment rationality.

Other exponents of the Frankfurt School add important insights regarding the value content of science as it becomes further linked to technology. Marcuse (1941/1982) defines technology not as productive tools (technics), but as a social process which shapes consciousness. As individuals come to identify with both technics and this structuring social process, they succumb to the fetish of technique – the willing hierarchichal coordination of activities for maximum expedience. Echoing Weber, Marcuse argues that under the sway of ‘technological rationality’ man values efficiency above all, and regards individualistic and critical values as obstacles to efficient production. But while Weber confined this value to specialities like engineering, Marcuse sees the notion of efficiency pervading throughout human social life. We increasingly view the world through the lens of what Lewis Mumford calls ‘matter-of-factness’ – in which:

“Individual distinctions in the aptitude, insight and knowledge are transformed into different quanta of skill and training, to be coordinated at any time in the common framework of standardized performances.” (Marcuse, 1941/1982, p. 142)

Marcuse is ambivalent about technics – machines and productive tools themselves. He believes these can serve the ends of liberation or exploitation.  

256 But his view of technology and

255 “The technical process, to which the subject has been reified after the eradication of that process [thought] from consciousness is free from…meaning altogether, since reason itself has become merely an aid to the all-encompassing economic apparatus.”

256 Marcuse develops his position on the liberating potential of modern technics in *Eros and Civilization* (1966/1955). There, he argues that technical progress has enabled mankind to overcome scarcity and the need for toil. Thus, he sees the continued compulsion to work as
technological rationality is far more pessimistic, as he believes the ideology of efficiency is particularly well-suited to absorbing critical protest and thus to preserving the existing order. This theme dominates his later classic work, *One-Dimensional Man* (Marcuse, 1964/1991). There he argues that the domination of nature and society are reinforced by a productive technological apparatus which has dispensed with metaphysics and become totalitarian (Marcuse, 1964/1991, pp. xlvi-xlvii). The Frankfurt School critique of technology thus relates the idea of value-neutral science to the maintenance of the exploitative productive relations of capitalism. In the evolution of this analysis, science also emerges as an relatively autonomous cultural force which shapes consciousness.

French social theory took a different path in understanding the politics inherent in science. Poststructuralism—a label used to refer to many primarily French social theorists of the 1960s and 1970s—offers its own critique of the knowledge produced by science. But unlike the Frankfurt School or phenomenology, poststructuralism do not offer an alternative grounding for pursuing scientific truth. Some of the best known works in this tradition come from Michel Foucault, and perhaps the most appropriate for the present purpose is *The Order of Things*. In his ‘archaeology’ of the human sciences—which focuses on the unconscious elements which shape the production of scientific knowledge—Foucault posits science and scientists as subject to surplus repression and thus believes that it is possible to reduce working time to its necessary minimum.

Poststructuralism of course must be a reaction to the tradition of structuralism. The latter is closely associated with the work of Claude Levi-Strauss who argues that there exist “vertical structures,” social formations which transcend space and time. These vertical structures are described in terms of Levi-Strauss’ famous binaries like ‘the raw and the cooked,’ and ‘the sacred and the profane’ (Levi-Strauss, 1963). Poststructuralists reject such transhistorical concepts and argue that all ordering concepts should be understood as historically constructed.

“…it is rather an inquiry whose aim is to rediscover on what basis knowledge and theory became possible; within what space of order knowledge was constituted; on the basis of what
forces beyond their control. Each historical period is determined by certain ‘epistemes’ which condition how knowledge is conceived and thus shape scientific discourse (Foucault, The Order of Things, 1994/1970). This archaeological approach also characterizes Foucault’s examinations of medicine, prisons and sexuality. In tracing each of these social formations, he emphasizes historical changes (both in the traditional and Marxist senses of that term) in the associated discourse: the language used to describe and delimit madness, crime, and sex.

An overarching theme of these diverse studies is Foucault’s concept of a ‘regime of truth’:

“[A] form of power which applies itself to immediate everyday life which categorizes the individual marks him by his own individuality, attaches him to his own identity, imposes a law of truth on him which he must recognize…” (Foucault, 1982, p. 781)

This concept is the critical link between power and the knowledge produced by science.

Although Foucault doesn’t regard power as residing with any specific person or institution, he emphasizes that modes of knowledge are constitutive of the forms of surveillance and domination This makes realltively clear the political nature of science – the prevailing contemporary mode of knowledge. Furthermore, the larger point of his work – that knowledge

historical a priori…sciences could be established…rationalities be formed , only, perhaps to dissolve and vanish soon afterward.” (Foucault, The Order of Things, 1970/1994, pp. xxi-xxii)

Foucault’s approach to the history of science – with its emphasis on discontinuity and contingency – comes from the influence of Gaston Bachelard through Georges Canguilhem. Bachelard developed the concept of the ‘epistemological break’ which is central to Foucault’s histories. Foucault’s early focus on medicine and his attention to discourse seem to be a product of Canguilhem’s influence.

Notably, Foucault looks at 19th century schooling and testing as evidence of this point: “Similarly [to the judicial penalty] the school became a sort of apparatus of uninterrupted examination… [it became] a perpetual comparison of each and all that made it possible both to measure and to judge…the examination is at the center of the procedures that constitute the individual as effect and object of power, as effect and object of knowledge.” (Foucault, 1977/1995, pp. 186-192)
production is not set apart from its social context – entails the certainty that science cannot be disinterested.

Louis Althusser, a contemporary of Foucault similarly influenced by Bachelard’s discontinuous history of the sciences, is much clearer about the relation between science and power. He argues that in moments of crisis for the sciences, the ‘spontaneous philosophy of of the scientists’ is revealed (Althusser, 1990/2012, pp. 115-116). This spontaneous philosophy is present, though dormant, in non-crisis periods as well, though most scientists would not recognize it as such. Like Foucault, Althusser regards the spontaneous philosophies of scientists as historically contingent and as consitutive of the limits of science (p. 128). But he goes further and identifies them, in the last analysis, with the dominant practical ideology of the historical period:

“…but they are used to furnish arguments or guarantees for extra-scientific values that the philosophies in question objectively serve through their own practice, their 'questions' and their 'theories'. These 'values' pertain to practical ideologies, which play their own role in the social cohesion and social conflicts of class societies.” (Althusser, 1990/2012, p. 130).

Althusser thus clarifies the relation of science and politics by arguing the philosophy which underwrites science at a historical moment is determined by the dominant ideology.\(^{261}\) Bringing him closer to the Frankfurt School tradition, he would regard science (and for that matter,

\(^{261}\) Similar to Husserl, who distinguishes between the natural and human sciences, Althusser sees values and ideology as more important in the latter (which he argues lack a true subject matter): “What we might call scientific ideologies and philosophical ideologies assume an extreme importance in the domain of the human sciences. Not only do these ideologies exist and have great importance in our world, but they directly govern the scientific practices of the human sciences. They take the place of theory in the human sciences.” (Althusser, 1990/2012, p. 92)
education) as one of the ideological state apparauses which shapes individual and social consciousness.262

I am left with the question of how science operates as an agent of power in shaping consciousness. The criticisms of value-free science above – particularly Marcuse and Althusser – have hinted at an answer; once science becomes the primary arbiter of knowledge in society, it becomes hostile to other epistemological forms. But the works of Paul Feyerabend and Stanley Aronowitz provide much more clarity on this question by focusing on the question of scientific method.

Feyerabend is most recognized for his ‘anarchic’ account of scientific progress in Against Method (Feyerabend, 1975/2010). He argues against scientific exceptionalism, demonstrating that science does not proceed according to a single pre-determined set of rules, nor is it guided exclusively or even primarily by rigorous observation. Rather, he contends that science, like other cultural and ideological formations, makes use of all means at its disposal to advance its propositions. Scientists also suspend what many regard to be the key tenets of the enterprise – falsifiability of theory and grounding in observation.263 His account shows that scientific progress would not have been possible without the eclecticism. Feyerabend ultimately views

262 Paul Feyerabend – discussed directly below – contends that Althusser accepts the ideology of science, as Althusser accepted the Marxist dialectical method as a positive science (Feyerabend, 1975, p. 177, see also Aronowitz, 1988, p. 180).
263 Feyerabend makes case studies of key moments regarded as points of great scientific progress (the Copernican Revolution, Einstein’s relativity, etc.). For example, he argues that Galileo helped overturn the geocentric model more through his ability to persuade than through his data or logical theories. As such, Against Method takes a different approach to the same data treated by Thomas Kuhn in his Structure of Scientific Revolutions.
science as not too dissimilar from myth or magic in its methods – an interesting point which we will consider again below.

Feyerabend’s view of science as far from value-free is implicit in this argument; if scientific methods are not distinct from other methods, then they are equally subject to political and ideological biases. Indeed, he argues that methodological choices are based on ethical decisions (Feyerabend, 1965, p. 219). But his work is much more specific on the point of how science actually operates as power – and this point focuses on methodology:

According to the fairy-tale the success of science is the result of a subtle, but carefully balanced combination of inventiveness and control. Scientists have ideas. And they have special methods for improving ideas. The theories of science have passed the test of method. They give a better account of the world than ideas which have not passed the test (Feyerabend, ‘Science.’ The myth and its role in Society, 1975, p. 167).

This special position afforded to scientific methods is important. If scientific methods are perceived as objective, then science – afforded the unique position of being value-free – thus becomes an objective measure of all ideologies (p. 169). Widespread belief in the objectivity of scientific methods leads to the use of scientific knowledge as the basis of important political decisions without recourse to democracy – since science produces untainted facts rather than ideologically-driven opinions.

Stanley Aronowitz (1988) brings together many of the ideas discussed above to synthesize an argument about the political force of science in modern society. With Lukacs and the Frankfurt School, Aronowitz contends that the rationalist model of science is linked to the domination of nature and of man. Similar to Feyerabend’s pluralistic epistemology, he begins with the premise that “the power of science consists, in the first place, in its conflation of knowledge with truth” (p. vii).
Making the case that scientific discourse has become hegemonic, Aronowitz shows that even critical theorists of science – for example Althusser and Marcuse – see scientific methods and their attendant technologies operating beyond social determination. They thus regard science as potentially liberating, but for their current use in the service of domination (pp. 165-180). This exclusion of scientific method from scrutiny is in part because intellectuals share what Alvin Gouldner calls a ‘culture of critical discourse.’ Aronowitz argues that this critical discourse is consciously and vigorously distinguished from ‘common sense’ by professional intellectuals.\textsuperscript{264}

The further and more substantial point is that “as scientific discourse permeates the state and civil society, scientific culture spills over beyond the laboratory,” affecting decision-making processes in all spheres of social life (p. 9). To the extent that scientific discourse – specifically its emphasis on objective methods – becomes hegemonic, non-scientists generally don’t question the knowledge claims made by scientists. These claims become truth itself. Moreover, approximations of the scientific ideal, controlled experimentation and mathematical calculation, become the standard of practice for political and industrial leaders.\textsuperscript{265} Finally Aronowitz argues that the remaining bastions of critical science – those portions of academia where the research programs are not in line with normative scientific activity (p. 320) and don’t indoctrinate new

\textsuperscript{264} Gouldner (1979) sees the role of this class as yet to be determined. He sees intellectuals as having revolutionary potential (as Althusser regarded science and Marcuse regarded technology), but still morally ambiguous and internally divided between those technically and humanistically-oriented. Steven Brint (1996) explores the internal divisions among ‘experts’ and argues that intellectuals are not so politically engaged and that many are in quite politically conservative.\textsuperscript{265} In this regard, Aronowitz also discusses government and foundation funding of so-called independent researchers to study/validate preselected policy options. But he also insists that these research mandates themselves are evidence of the independent power of scientific ideology – as political and economic powers feel compelled to cloak their decisions in the veneer of scientific method (pp. 326-336).
researchers – science all but completely merges with power as a single version of science becomes the arbiter of natural and social truth.

On Mathiness and Policy

Most if not all the scholars (positive and critical) cited above discuss the role of mathematics in the practice and ideology science. Comte regarded mathematics as first among the sciences and a methodological precursor to positive science. Husserl wrote on the ‘mathematization of nature’ in his phenomenology. Althusser regards mathematics not as a tool of the sciences, but as an active participant in their constitution. And given the depth of Value-Added’s mathematical content, one must account for the particular place of math in the politics and practice of science, and particularly in economics – from which Value-Added hails.

Economics leads the social sciences in its capacity to influence policy. The president maintains a council of economic advisers, not sociological advisers. The Congressional Budget Office and the Federal Reserve are two government institutions dominated by economists. Indeed, we saw above how an economistic mindset came to dominate federal education policy beginning in the 1970s; that trend has not abated. Thus, the role of math in the internal dynamics of economics is particularly important to understand.

It is well-known that economics is dominated more than most social sciences by sophisticated mathematical modeling. A casual glance at a mainstream economics journal will show that even what economists call theory is essentially equation-building.266 Economics did not begin this

---

266 As Levinowitz (2016) notes, Deirdre McCloskey and Robert Nelson both see mathematics used in economic theory as a way to stake a claim to scientific status. As I’ve pointed out above, Raj Chetty (2013) has gone out of his way to emphasize the objective scientific credentials of his discipline.
way, but it was pervasive enough by the early 20th century that no less than John Maynard Keynes wrote in his general economic theory that:

“Too large a proportion of recent ‘mathematical’ economics are merely concoctions, as imprecise as the initial assumptions they rest on, which allow the author to lose sight of the complexities and interdependencies of the real world in a maze of pretentious and unhelpful symbols.” (Keynes, 1936, p. 297)

But recent work on the influence of mathematics in economic theory regards this overemphasis on modeling as politically suspect and counterproductive to economics as a discipline. Paul Romer, an economist who has made notable contributions to the theory of economic growth, recently argued that his sub-field is being dominated by ‘mathiness’ which he defines as:

“[using] a mixture of words and symbols, but instead of making tight links [between the two] it leaves ample room for slippage between statements in natural versus formal language and between statements with theoretical as opposed to empirical content.” (Romer, 2015, p. 89)

He distinguishes mathiness from mathematical theory, specifying that the latter makes tight links between mathematical symbols and the language used to describe them. Such also allows theory to be closely connected to the empirical evidence used to support it. Romer cites Gary Becker’s human capital theory as an example of strong mathematical theory and good science. He contends that unfortunately, economics also has a tendency toward ‘academic politics;’ in the context of this research, we would call this ideology. The problem for Romer is that mathiness allows ideology to “masquerade as science” (p. 89), since the proliferation of mathematical economics makes it hard to separate the proverbial ‘wheat from the chaff.’ He worries that prolonged exposure to poorly-specified mathematical models without solid foundations will turn

---

267 What Romer is doing in this argument is a well-established tradition. Althusser (1990/2012) points out that scientists react to crisis in their discipline by trying to root out what they see as poor practices.
economic theory into something like card tricks, where everyone expects sleight-of-hand but nobody takes it seriously (p. 93).

Other economists have taken the sleight-of-hand metaphor further when commenting on mathiness. Whereas Romer’s mathiness refers specifically to macroeconomic growth theory, Paul Pfleiderer (2014) makes a similar case for finance and economics more generally. He posits a distinction between ‘bookshelf’ and ‘real-world’ economic models, noting that while all economic models are unrealistic to a degree, bookshelf models are not intended to correspond to reality and should be regarded as similar to thought experiments. The practical distinction between the two is that bookshelf models contain assumptions that seem impossible in the real world. Take as two examples: ‘there are no information asymmetries between buyers and sellers’ or more to our point ‘teacher effects on student academic performance continue undiminished forever.’ Pfleiderer argues that when such bookshelf models are advanced as support for policy prescriptions in the real world, they become ‘chameleons.’ The problem is that once economists are criticized for the dubious assumptions made by such models, they retreat to the position that their models are not intended to correspond to reality.268 One of the primary ways that chameleon economic theories succeed in penetrating into policy discussions is that they are often cloaked in complex mathematics:

“A simple model with dubious assumptions that drives the results is likely to be recognized for what it is. A much more complex model with many assumptions including the aforementioned dubious assumptions, is more difficult to evaluate.” (Pfleiderer, 2014, p. 28)

268 Eric Hanushek’s testimony in Vergara v. California is exemplary of such a chameleon posture. When challenged about the application of VA to the real world, Hanushek replies that nobody ever suggested that Value-Added Assessment be applied as the sole metric of teacher quality (other than he for his entire academic career).
Ultimately, Pfleiderer argues that all theoretical models – plausible, absurd, and chameleon – must be evaluated on the strength of their assumptions, and that defenses such as models being ‘true pending empirical falsification’ or that some dubious assumption ‘is standard in the literature’ have to be challenged.

Where does all this mathematical economics lead us? Alan Levinowitz (2016) compares present-day economists’ fetishization of mathematical models with similar valorization of mathematically-sophisticated astrology in ancient imperial China. Like modern economists, Chinese astronomers possessed enormous influence in the imperial court, which believed that ever-more accurate calculation of celestial motion (Li) would provide a guide for policy decisions; but also like modern economics, many basic assumptions of astrology are dubious – something that no amount of sophisticated calculation can fix. Nonetheless, the allure of their models gave the Li-calculating astrologers access to power and quite large sums of research funding from the Emperor (Levinowitz, 2016, p. 5). Levinowitz thus argue that, then as now, “when mathematical theory is the ultimate arbiter of truth, it becomes difficult to see the difference between science and pseudoscience.”

While I would stop short of calling economics a pseudoscience, it is safe to say that the high-level mathematics it employs grants an imprimatur of objectivity which allows its assumptions to go relatively unquestioned. We saw this at numerous points in the debate over value-added assessment. It was precisely this assumptions-versus-data contest that we saw in Vergara. Jesse Rothstein’s work on “Teacher Quality Policy When Supply Matters” – which was the substance of his expert testimony – tested the assumptions of Value-Added-based teacher dismissal policies against the real-world characteristics of the teacher labor market. The plaintiffs lawyers protested that all this was specualtion, arguing that Thomas Kane and Raj Chetty had presented “actual
data” to support their claims, without mentioning that these data – via the modeling strategies employed – are products of their underlying assumptions. What seems clear from the trial verdict is that the data presented by Chetty, Kane and Hanushek were enough to ‘shock the conscience’ of the presiding judge and those of the media outlets, while Rothstein’s critiques of model assumptions could not moderate the shock.

Bringing it all together – Value-Added and moments of “Methodological Politics”

We’ve outlined two broad positions above; the first is that science and technology constitute a value-free enterprises capable of conveying relatively unmediated truths about nature and society. This position is the implicit operating assumption of most mainstream social science. Various critical perspectives share the view that science has inherent ideological content which is primarily associated with technical efficiency and top-down control of natural and social forces.

Despite the singularly humanistic enterprise that constitutes its object, much education research shares the implicit assumption of value-neutrality. In his work on the politics of charter school research, Henig (2008) argues that even though this body of research was put to overtly political uses by both charter school proponents and their opponents in the teachers’ unions, the processes of data collection and analysis remains shielded from political influence.²⁶⁹ This claim is tenuous even in Henig’s own analysis; he concedes that private philanthropies – which play a large and growing role in funding education research – are very specific about the questions sponsored research asks, the modes of inquiry by which they are answered, and the ways research findings

²⁶⁹ This is a version of Reichenbach’s (1938/1977) distinction between the “context of discovery” and “context of justification.” Henig argues that the discovery of objective knowledge about charter schools remains a value-neutral process even if the reasons for first inquiring about them were infused with politics.
are presented. If funders – most of whom have very overt ideological positions – determine what is asked, how it is answered, and how the answers are reported, what of this process is value-free?

Regarding federally-funded education research, Margaret Eisenhart (2005) contends that ‘scientific’ research stands apart from moral and political questions. Responding to criticism of a 2002 National Research Council Report, Scientific Research in Education – on which she collaborated and which defined scientific research in primarily quantitative and experimental terms – Eisenhart says political and moral questions are best left to ‘philosophers’ and ‘critical theorists’ (p. 54) since science focuses on gathering empirical evidence about ‘what works.’ This position is similarly poorly-founded; even the NRC definition of scientific research is admittedly determined by the opinions of congressional leaders, and the political resonance of (quantitative and experimental) research (2005, pp. 52-53). Again, if funding sources private and public determine a validity hierarchy for research methods (quantitative/experimental vs. qualitative/descriptive) and specify the nature of research questions (emphasizing those which create interventions on pre-determined outcomes) at the outset, education research can hardly be described as standing apart from value judgments.

The persistent claims to value-neutrality perhaps represent the aspirations of social scientists powerfully articulated by Max Weber, even though Weber himself understood this as an ideal and knew that social science could never stand apart from value judgments. But, as the critical tradition outlined above suggests, ‘objective’ research in education seems to carry and conceal its

---

270 Baez and Boyles (2009) argue that the language of ‘scientific research in education’ in the No Child Left Behind Act, which greatly influences the NRC report, is an instance of political interests using science to subvert politics.
own particular value set. Indeed, the recent fashion for discovering “what works” in education implies that research is only supposed to find more efficient means of producing already-accepted outputs which are themselves determined by the political process. This was precisely the danger alluded to by John Dewey in *Education as Engineering* (2009) and by Walter Lippman in his critiques of intelligence testing discussed in detail in chapter 3 above.

But the more interesting question that the above study sheds light on is how the value set of objective research plays out in moments of technological innovation. Critical scholarship on science indicates that allegedly value-neutral, and specifically mathematically sophisticated research is vested with an imprimatur of objectivity. We’ve seen that knowledge stated in numbers and equations receives special treatment in literal courts – Frederick W. Taylor’s congressional testimony in 1912 or the *Vergara v. California* trial in 2014 – and in the ‘court of public opinion’ as we saw with the Boston School Committee report in 1845. Today, even politically progressive forces in education argue for more comprehensive data systems and research-proven interventions, believing that they provide unmediated truth (Education Opportunity Network, 2013).

This fetishization of mathematical methods and quantitative data in general seems to be amplified by novel techniques; innovations are accompanied by millenialist hopes that education will be once and for all be rendered productive, efficient, accountable, and fundamentally equal for all students. We saw this sort of hope with the emergence of large-scale experiments and systems analysis in the 1970s and we see it again today. In such moments, we forget Weber’s assertion that policy cannot be decided by simple recourse to facts; cultural values remain essential to establishing social priorities.
Thus, while Raymond Callahan (1962) and Jal Mehta (2013) ascribe the success of top-down education reformers to the relative weakness of teachers’ organizations, these successes are equally well-explained by the unwarranted moral and political power afforded to science and technology at moments of innovation. As Morozov (2013) and Aronowitz (1988) suggest, the logic of domination is smuggled in and democratic politics are crowded out by the belief that difficult educational problems can be solved with seemingly innocuous technical improvements to the mechanisms of control. But we see that these technical changes are always accompanied by new pushes for administrative centralization. Whether it is the move for professionalized superintendency in the 19th century or the gutting of teacher tenure protection in the 21st century, technology is politics carried out by other means.

Further, the examination of the Vergara trial in chapter five adds another micro-level layer to the issue of objectivity in science and policy that complements this idea of technology as politics by other means. The expert witnesses in that trial would all claim to be doing value-free science rooted in objective empirical methods. But given the very political implications of the case – the elimination of teacher tenure protections – the experts’ opinions vis-à-vis this consequence, are expressed in peculiar forms.

Take a few examples. Raj Chetty’s testimony for the plaintiffs features an abundance of sports analogies: baseball statistics and Kobe Bryant’s height and why coaching athletes still matters. Sports metaphors allow Chetty to talk about the quantitative accuracy of a technique like Value-Added in terms set apart from its real consequences in the very fraught terrain of California
public schools. Tom Kane’s discussion of the test score impacts of ‘teacher deselection’
similarly veils the true potential effects of VA-driven teacher evaluations – teachers will be fired
and lose their livelihoods based on unreliable and perhaps invalid measures of their abilities. In
his testimony for the defense, Jesse Rothstein discusses abstract assumptions about teacher
supply at length, critiques Chetty’s unorthodox definition of forecast bias, and underscores the
high rate of attrition in the Measures of Effective Teaching Project random assignment study.
These are strong methodological arguments, but they fail to capture what’s at stake. In order to
refute these criticisms, Eric Hanushek comes full-circle to the origins of this technique and
recasts VA as a research method used by economists, not as a policy tool with consequences.
Value-added as a decision-making policy is held at arm’s length despite the fact that such was
the very substance of the Vergara trial.

These exchanges point to what I regard as a ‘methodological politics.’ Chetty, Kane, Rothstein
and Hanushek each have very clear positions on the appropriateness of VA as a tool for high-
stakes teacher evaluations. Yet, their testimonies become a back-and-forth of statistical minutae;
instead of teachers’ livelihoods, experts debate standard error calculations and the definition of
forecast bias. This may be an artifact of the highly scripted nature of expert testimony, and of the
particularities of economics as a discipline (described briefly above), but the exaggerated nature
of these exchanges is a useful way of looking at science as a political force. In essence, experts
who all purport to be pursuing value-neutral truth attempt to best each other with statistical
sophistication and force their opponents to accept their value judgments as the necessary
consequence of their ‘objective’ data and methods. If one accepts the methods by which Chetty

271 Promotional materials for value-added assessment go back to the technique’s origins in
agricultural research and talk about VA in terms of trees growing.
empirically adduced *Long Term Impacts*, or the ‘gold standard’ of experimental design employed in *Measures of Effective Teaching*, one must therefore accept the implication that teachers ought to be evaluated by student test score gains. This style of discourse obscures the implications of scientific knowledge in favor of abstract technical discussion. But politics are only obscured, not altogether absent; the political contents remain intact, as does the appearance that science stands apart from such matters. Thus with Marcuse, we might argue that:

“The analysis is ‘locked;’ the range of judgment is confined within a context of facts which excludes judging the context in which the facts are made, man-made, and in which their meaning, function and development are determined.” (Marcuse, 1964/1991, pp. 115-116)

Placing Value-Added Assessment in context of the history of innovations in educational assessment provides perspective on how science manifests as power in determining policy. At the larger level, science proceeds as a form of politics to the extent that new technologies are perceived as merely technical innovations free of inherent biases. Yet, historical examination shows that the use of these techniques is always accompanied by an attendant push for increasingly centralized control over schooling. The perceived neutrality of scientific methods and the tools they produce allows this admisitrative control to be smuggled in. On a smaller scale, debates over policy are displaced into debates over technique, constituting what I have called a methodological politics. In the case of value-added, the real stakes of evaluation policies are submerged beneath expert opinions on the precision and validity of measurement techniques. More importantly, the distinction between fact and value is blurred, and experts who engage in such discourse seem caught in a bind. Their commitment to the ideal – or the appearance – of objectivity colors the very political consequences of their work, resulting in a contest over methods and data rather than interpretations and implications.
Something else gets lost in abstract discussion of teacher evaluation policy – education. Schooling is a structured interaction between teachers and students which is supposed to educate the latter. Having worked with K-12 students throughout the course of this research, it behooves me to keep them in mind; at some points I have done so. Chapter one assessed how the workings of schools and classrooms would have to be transformed to meet value-added model assumptions, and placed this possibility in context of the necessarily contentious relationship between teachers and students. There I argued that overemphasis of tests and their results would vitiate the important moral content of education. Similarly, the history of assessment technology explored in chapter three repeatedly invoked the hazards of measurement-driven instruction for both teachers and students. These are important topics, but insofar as this analysis has been confined to schooling as a process rather than education as its goal, it is necessary but not sufficient.

Education is not reducible to schooling; it is rather the socially agreed-upon goal which schooling – concurrent with other social institutions (notably the family, media, and popular culture) – is supposed to achieve. Education is successful to the extent that it reproduces agreed upon values, knowledge and capacities among future generations. Understanding the impact of evaluation technology on education as distinct from its impact on schooling requires defining these shared goals – what is to be taught and learned.

A comprehensive statement on K-12 curriculum is far beyond the scope of this project, but some general outlines can be traced. Curriculum has been and continues to be a ‘third rail’ in US politics, particularly when the time comes to discuss evolutionary biology, sex education, and certain aspects of American history. In addition, the notion of a national curriculum runs up against the idea of ‘states rights,’ making it a wedge issue for conservatives. The ‘Common
Core’ standards – the most recent attempt at K-12 curriculum policy – bear the imprint of both of these forces; these standards primarily focus on mathematics and reading, and many states (especially in the south) either banned the standards altogether or never adopted them in the first place.

In the absence of any sustained explicit negotiations, curriculum is defined by assessments and thus by textbook publishers and test-makers. While many industrialized countries use comprehensive content examinations as the capstone of students’ compulsory education, the closest US equivalents are the SAT and ACT college entrance exams. Even students who apply to open-admissions and community colleges are faced with some version of the SAT/ACT. Despite numerous revisions intended to align these exams with curriculum, they are still a hybrid of basic literacy and numeracy and aptitude testing. These tests are the benchmark of US educational assessment, and thus of the curriculum itself. Teacher evaluation systems which emphasize test results – forcing teachers who have in the past served as a buffer zone between students and the tests – only exacerbate this test-defined experience of schooling.

What does test-defined schooling imply for education in terms of values, knowledge, and capacities? Such exams place a premium on reductive thinking and short-cuts (Hoffman, The Tyranny of Testing, 1964/2003). They reward students who can identify the least rigorous path to the test-maker’s wanted answer and who can temporarily suspend critical thought when faced with two or more ‘possibly correct’ choices. Even the allegedly free-response essay portions of these exams – which are optional in the case of the SAT and ACT – encourage writing to a template rather than crafting original ideas. If value-added evaluation systems disproportionately reward teachers who foster these skills in their students, then education is reduced to test-preparation. Rather than fostering critical thought and innovation, such instruction thwarts those
goals by rewarding their opposite. Moreover, to the extent that high-stakes decisions are based on value-added measures, teachers are prevented from acting autonomously as professionals in the classroom and are instead forced to work within the constraints imposed by the tests.

Why is educator professionalism important? The production of educated students is not a one-size-fits-all process; and it contains both measureable and unmeasureable goals. In this process, educators need room to adapt to the students they engage with. Early proponents of standardized testing understood this and hence emphasized the use of exams as tools rather than as curriculum itself (Chauncey & Dobbin, 1963). But this message fell by the wayside as policymakers clamored for more data by which to judge students, schools and teachers. Value-added is perhaps the culmination of this trajectory; it reduces the whole of a complex relational process to the residual term in a loosely-theorized and poorly-specified mathematical equation.

Ultimately, what I take away from this research is that science and technology can blind us to the very complex and contingent realities of vital social processes such as education. Insofar as technical innovations pose enticing and easy solutions to seemingly intractable problems, they can also temporarily corrupt the politically difficult but necessary process of devising more lasting solutions. Schooling requires more than a series of poorly-informed panics and technological panaceas; it needs sustained investment and a level of study whose nuance matches its level of complexity. Simply put, to fulfill its vital social mission, education requires more than just endless measurement.
Postscript on Value-Added and the return of politics

Writing takes time, and history continues unabated. But in this case it seems history has helped validate my thinking in two ways which I will mention briefly. My historical analysis in chapter three suggested that after periods of fervor for methodological innovations, they tend to fall out of favor. Though it is still uncertain, value-added seems to be on that path. In New York, Governor Mario Cuomo – who was a staunch supporter of value-added assessments – has since backed off the mandate for teachers to be evaluated on the basis of test scores – even potentially reducing their proportional role to zero (Taylor K., 2015). In California, the Vergara ruling was overturned by an appeals court on April 18th, 2016 - although Students Matter plans to take the case to the Supreme Court.

What is more important is why this is happening. One important factor is the opt-out movement, which encourages parents to refuse to have their children sit for local, state and national-level standardized tests. Change the Stakes – a New York City-based opt out group – lists standardized tests’ role in evaluating teachers as among the primary reasons to resist testing. Supporters argue that the perverse incentives created by test-based evaluation hurt instruction at both the classroom and school level by encouraging teachers to reallocate time and discouraging teacher cooperation. A sign of the appeal of opt-out, 20 percent of New York State students in grades 3 through 8 did not sit for testing (Harris E., 2015). Growing opt-out numbers are also evident in Washington, New Jersey, Pennsylvania and Illinois (Stokes, 2015).

Opt-out, in light of the above research, seems a logical response to the crowding out of democratic politics by technology in the field of education. Policymakers see the world at a particular level and want aggregate evidence at that level to make decisions (Scott, 1998). But, as
Feyerabend makes clear, the reliance on the so-called objective evidence in science is fundamentally undemocratic. It is presented as unassailable truth rather than as knowledge produced by human beings with particular values of their own, and rejects all other knowledge claims as ideological (Feyerabend, 1975). Thus, when opt-outers declare that “my child is not a number” and seek to revitalize the interactions of the classroom – including their own participation – they articulate a fundamentally democratic response to a top-down methodological politics epitomized by value-added teacher evaluations. And in targeting the supply of data on which these evaluations subsist, opt-out demonstrates its recognition of the role of science and its methods as critical mechanisms of power in society. Understanding full well that my own science is not value-free, I hope they succeed.
Bibliography


Elliot, E. (1915). How Shall the Merit of Teachers be Tested and Recorded. Educational Administration and Supervision, 1, 291-299.


Murphy, K. (2014, June 13). Key witness says he was misquoted. LA Daily News.


Zeichner, K. (2013, August 27). How the public is deceived about 'highly qualified teachers'. Retrieved August 5, 2015, from The Answer Sheet: