THE LIMITS OF LEGIBILITY: WHY ACCOUNTABILITY-BASED EDUCATION REFORMS HAVE NOT BEEN A PANACEA

By

DAVID HERSH

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Planning and Public Policy

Written under the direction of

Radha Jagannathan

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

January 2020

ABSTRACT OF THE DISSERTATION

The Limits of Legibility: Why Accountability-Based Education Reforms Have Not Been a Panacea

By DAVID M. HERSH

Dissertation Director

Radha Jagannathan

For over a century, reformers have sought to fix public education with an increasingly intense focus on individual accountability. The round of reforms beginning with the No Child Left Behind Act (NCLB) in 2003 and culminating with state-implementation of Race to the Top (RTT) initiatives over a decade later was arguably the biggest bet yet on accountability. Neither past reforms nor this latest effort has achieved the lofty goal of dramatically changing outcomes for America's students. This dissertation aims to explain why, applying a framework derived from similar failed efforts to govern complex natural and social processes in other fields to case studies of three school districts in New Jersey. Using this framework, I argue that accountability efforts as implemented in education fail because they meet three conditions: they over-simplify the highly complex social process of educating children, they shift expertise from educators with local knowledge to distant "objective" outsiders and they ignore contextual differences between students, schools and districts. These three conditions lead to failure because they trigger at least four mechanisms that get in the way of improving outcomes. They lead to policies based on

mischaracterizations of the problem that engender resistance and undermine conditions necessary for successful instruction. Given the difficulty of designing a summative accountability scheme that does not meet these three conditions of failure, a better path may require prioritizing formative efforts over summative accountability.

Dedication

For Mom,

whose dedication to her students

continues to shape my perspective.

And for Samara,

whose support made this possible.

Acknowledgements

I am endlessly grateful to the many people whose assistance was necessary for me to complete this dissertation. No one played a bigger role than my advisor, Radha Jagannathan, without whom I would be among the many whose academic career ends ABD. In addition to reading hundreds of pages, the majority filled with far-less-than-riveting prose, Dr. Jagannathan's pragmatism and encouragement ensured I continued to make progress. Her influence shaped my education and career as much as it shaped this dissertation.

I am likewise grateful to my other three committee members, Julia Sass Rubin, Hal Salzman and Drew Gitomer. Throughout this project and my more general studies during my time at Rutgers, each provided invaluable insights and support that inform the content, tone and conclusions in this dissertation.

No less critical were the teachers, principals and district administrators who generously gave their limited time to speak to me about their experiences. Their candor created a vivid image of their working lives. Their expertise continues to inform my vision of education policy.

A larger group of professors, administrators and colleagues at Rutgers influenced or supported me in this work. With apologies to those whose names I will undoubtedly miss – a reflection of just how long this took and my increasingly unreliable recollection – my partners in this work included Angie Oberg, Ryan Good, Lee Polonsky, Morgan Campbell, Janice Fine, Bob Lake, Stuart Shapiro, Gabriella Carolini, Steve Weston and Courtney Culler.

I was also influenced by discussions and shared experiences with professional colleagues in the district in which I worked. I hope they will forgive the absence of their names, which, if listed, would undoubtedly make obvious the identity of one of the subject districts of this dissertation.

Finally, thank you to the family and friends who indulged with good-spirited mocking my seemingly endless career as a student. Most importantly, thank you to my parents, who did not waiver in their support despite my apparent efforts to avoid adulthood, and to my wife, Samara, who gave up countless weekends and breaks from taking care of our son so I could finish this.

Table of Contents

List of tables

List of Illustrations

1 Introduction

Perhaps more than any other public service, public education is subject to almost constant calls for reform. For more than a century, the answer to these calls has been greater accountability. The latest round of reforms, which for this dissertation I define as the test-based accountability reforms beginning with the No Child Left Behind Act (NCLB) in 2003 and culminating with state-implementation of Race to the Top (RTT) initiatives over a decade later, is arguably the biggest bet yet on accountability. This is despite the lack of evidence that accountability-driven reforms had resulted in dramatic gains in the past. Still, these latest efforts added another level of rigor to accountability, leveraging mandatory testing and longitudinal data systems to more closely link educators to the performance of their students. Perhaps the problem with past accountability efforts was simply that they did not go far enough?

That does not appear to be the case. Two of the most comprehensive reviews of teacher evaluation – the core of the latest round of reforms – find little evidence that it is resulting in meaningful improvements in student outcomes. One review sums up the findings of the Intensive Partnerships for Effective Teaching, an effort funded by the Bill & Melinda Gates Foundation (BMGF). The partnership included three traditional public school districts and four charter management organizations (CMOs) in California and lasted from 2009-10 through 2015-16. All of the districts had implemented measures of teaching effectiveness that included student growth as measured by standardized tests and a measure of teacher practice. The culminating report, "Improving Teaching Effectiveness: Final Report," sums up the results. Despite evaluating teachers with

measures of teaching effectiveness and using those evaluations in human resources (HR) decisions, the sites "did not achieve their goals for students." (Stecher, et. al. 2018). In a more global review that delved deep into the performance management literature, Rowan and Raudenbush's chapter in the Fifth Edition of the Handbook of Research on Teaching (2016) came to a similar conclusion. "The weight of the empirical evidence…suggests that using objective measures of teaching performance in consequential decisions… has not generally produced the large benefits expected by education reformers." (Rowan, B. & Raudenbush, S. 2016 p.1208).

While reviews of the latest accountability-driven reforms and their predecessors have identified some of the reasons for the failures in their specific instances, the consistency with which these efforts have failed to result in large gains begs for a more comprehensive explanation. This dissertation aims to provide such an explanation, applying a framework derived from failed efforts to govern complex natural and social processes in other fields to case studies of three school districts in New Jersey. This framework suggests that accountability efforts as implemented in education fail because they meet three conditions: they over-simplify the highly complex social process of educating children, they shift expertise from educators with local knowledge to distant "objective" outsiders and they ignore contextual differences between students, schools and districts. These three conditions in turn lead to failure because they trigger mechanisms that run counter to success. They mischaracterize the problem, engender resistance, undermine conditions necessary for successful instruction and are generally implemented poorly. Given the difficulty of designing an "objective" accountability

scheme that does not meet the three conditions of failure, a better path may require prioritizing formative efforts over summative accountability.

1.1 Education Reform Déjà vu

1.1.1 A long history of reform failures

In her 2012 book critiquing the latest education reforms, Diane Ravitch, formerly an architect of NCLB, refers to the historical "rise and fall of grand ideas that were promised as a sure cure…" for problems in public education. (Ravitch 2012, p.3). In this, she echoed earlier histories of education policy, particularly those of Callahan (1962) and Tyack (1974). A key theme is the persistent failure of prior efforts, well captured by Stephens (1967):

> "Every so often we adopt new approaches or new methodologies and place our reliance on new panaceas. At the very least we seem to chorus new slogans. Yet the academic growth within the classroom continues at about the same rate...[W]e would be making a great mistake in regarding the management of schools as similar to the process of constructing a building or operating a factory... We start, on the contrary, with a complex and ancient process, and we organize our efforts around what seeds, plants, and insects are likely to do anyway. When teachers and pupils foregather, some education may proceed even while the Superintendent disports himself in Atlantic City." (p.9-11).

Others have echoed this. Both Tyack's (1975) and Callahan's (1962) histories offer explanations for a century of failed accountability reforms. Meyers and Rowan (1978) frame their discussion of the institutionalization of education organizations against a history of ineffective accountability pushes. Grant (1989) and Chubb and Moe (1990) likewise frame their analysis against a history of reform failure.

With such a consistent failure narrative, a brief reflection on the definition of failure is warranted. As Stephens (1967) quote suggests, failure in this narrative does not mean the absence of growth but rather growth that does not accelerate enough with each

reform to satisfy public education's many stakeholders. To be sure, evidence suggests that, when measured by the traditional outcomes targeted by reformers, there has been a roughly 3-4% annual improvement. (Salzman, H. and Benderly, B.L. 2019). What needs explanation, then, is why progress has been so steady in the face of such frequent pushes for dramatic improvements (with correspondingly large resources devoted to the enterprise). This is the question this dissertation seeks to answer.[1]

The history of reform offers the beginnings for an explanation of the consistency of results: while reform is ever present, it always seems to have the same basic features. More than twenty years after Stephens wrote, Chubb and Moe (1990) referred to "a sense of déjà vu" in pushes for education reform, noting the similarities between both the drivers of then-current and past reforms and the basic structure of reform efforts. The motivations generally derived from fears about competitiveness. In turn, the solutions focused largely on increasing accountability – especially of individuals - through standardized testing. Reforms undervalued context, organizational characteristics, and the limits, challenges and implications of testing.

Perhaps no work better exemplifies the narrative that education reform is needed to stave off a disastrous failure to compete than "A Nation at Risk." (United States 1983).

[1] I acknowledge that there is a second question implicated by this frame that this dissertation will address only peripherally: Why is the progress that has been made so unsatisfactory to stakeholders that reform is always called for? This acknowledges, for example, that many stakeholders may have an interest in the failure narrative independent of whether or not reforms result in meaningful progress, reflecting, among other things, public educations complicated position as a political football serving purposes beyond any outcomes over which schools have control. Although this question is important, I focus on the results of reform because my interest is more in how reform plays out in schools than about what triggers the reforms in the first place. Thus I address the motivations of stakeholders only to the extent that it impacts the experience in the classroom (by, for example, delegitimizing teachers as professionals).

Created by then-Secretary of Education T.H. Bell in 1981, the National Center for Excellence in Education (NCEE) was charged with developing a comprehensive report on the state of the nation's education system. NCEE had 18 months to deliver its report, including findings and actionable recommendations. The first two sentences of A Nation at Risk make the need for reform clear: "Our Nation is at risk. Our once unchallenged preeminence in commerce, industry, science, and technological innovation is being overtaken by competitors throughout the world." (United States 1983, p.1). The changing job market, demanding every higher skills - particularly in what we now call STEM - in an increasingly international market, was one of the biggest reasons for the threat. The blame for our decline lay squarely at the feet of educational mediocrity. "The educational foundations of our society are presently being eroded by a rising tide of mediocrity that threatens our very future as a Nation and a people." (United States 1983, p.1).

A Nation at Risk identified teaching as only one of four aspects of the educational process in need of reform. It also laid the blame for teaching deficiencies at the feet of teacher supply, especially in key fields like mathematics and science, where shortages driven by low-salaries and poor preparatory programs led to a large share of unqualified teachers. There was no mention of insufficient incentives or a lack of accountability in the findings. Notably, there was no recommendation that teacher accountability be enhanced. Nevertheless, a key recommendation was that schools adopt "more rigorous and measurable standards" through standardized testing and that colleges and universities raise their requirements for admissions. These were, however, aimed at sorting students, not establishing the effectiveness of educators. (United States 1983). Thus, despite being a landmark in the public education debate and a lightning rod for critique, A Nation at

Risk to some degree diverged from a historical and subsequent obsession with educator accountability. Its needs assessment, however, was right in line with that history and the narrative under-girding the round of education reforms addressed here.

Tyack (1975) offers a comprehensive review of the history of education reform preceding A Nation at Risk. After tracing the development of the education system from the 1850's to the 1970s, Tyack blamed a reliance on individual accountability for the lack of results: “One of the chief reasons for the failure of educational reforms of the past has been precisely that they called for a change of philosophy or tactics on the part of the individual school employee rather than systemic change..." (Tyack 1975, p.10). Instead, he thought it "more important to correct...the system...than to scold its agents...” (Tyack 1975, p.10). Relatedly, he found that ignoring context was a problem as far back as the 19th century, noting that "schoolmen developed ideological and organizational consensus in their search for educational order, but heterogeneous values among urban populations and diffusion of power in school governance frequently complicated their task." (Tyack 1975, p.7).

Current policymakers might also be struck by the specifics of 19th century reforms and the surrounding debates. Standardized curricula and fears of teaching to the test date to the 1850s. Meritocracy based on testing was argued for in the 1870s. John Philbrick, a prominent New England principal and superintendent, argued that the future depended on better tests for evaluation in 1885. The late 1800s also saw pushback in the form of arguments about the inappropriateness of using test scores to compare schools and teachers. (Tyack 1975).

The similarities continued with the top-down reform campaigns of the 1890s to

the 1920s. Driven by an upper class that saw bureaucracy as a means of disinterested control, reforms were pushed through even though broader public sentiment was very different than the norms reformers used to justify the changes. The belief in the objectivity of scientific measures led to a growing role for testing. By the 1940s the use of IQ tests to categorize students became a self-fulfilling prophecy. "Objective" scientific measures produced genuine unforeseen consequences for black Americans. (Tyack 1975). This echoes a concern John Dewey raised for standardized tests in the 1920s, likening the sorting of students via tests to the creation of a scientifically legitimized caste-system. (Dewey 1922a). Through the 1940s, accountability was the mantra used to justify changes that largely ignored reality. (Tyack 1975).

In the 1940s and 1950s teacher shortages driven by demographic and structural factors became a significant concern, and some began to fear that reforms might be driving the best people out the system. The 1960s saw a renewed reform push driven by fears over competitiveness in education's "Sputnik moment." Owing to additional concerns about whether equality of input yielded equality of output, there was a big push for vouchers, performance contracts, decentralization, free schools, and alternative schools within the public system. By the 1970s, however, reformers were discouraged by the lack of results. Tyack was not surprised, noting "inadequate solutions [were] implicit in simplistic definitions of what constituted the problems." (Tyack 1975, p.8). This “Sputnick moment” driver of reform was part of Chubb’s and Moe’s (1990) "sense of deja vu” as the "Sputnik moment" was repeated amidst the economic problems of the 1980s, most notably in A Nation at Risk.

Callahan (1962) traces a similar history to that of Tyack, finding that by the 1960s

there had been over 100 years of developing measuring sticks. In particular, there were measures of teaching effectiveness before 1900, by 1910, tests were constructed and standardized and by 1912 there were new rating sheets for teachers based on scales from "objective" achievement tests in literacy and math. 1923 saw debates about "time-wasting, energy destroying statistical research" and the merits of tests for measuring efficiency. (Callahan 1962, p.122). Many of these changes were driven by the demands of the business community and an obsession with scientific management that had weak administrators scrambling to turn schools into factories. (Callahan 1962). More recently, Ingersoll (2003) noted that accountability has been a recurring theme since the turn of the century with "advocates argue[ing] that one solution to the problems in schools is to institute more external controls on teachers." (p.234-5).

Thus we see a consistent pattern in the history of education reforms. First, reform advocates frame reform as a necessary means of enhancing international competitiveness to stave off national disaster. Second, reforms focus on accountability based implicitly or explicitly on a highly simplified definition of the problem. Finally, reforms include technologies to facilitate accountability, often in the form of standardized tests and evaluation rubrics.

1.1.2 The latest round of reforms

Against this backdrop, the latest generation of reforms begins to look like déjà vu all over again. Announcing the passage of No Child Left Behind in 2002, then-President Bush laid out the principles that would "help guide our public school system for the next decades." In his words, the

> [*f*]*irst principle is accountability.* Every school has a job to do. And that's to teach the basics and teach them well. If we want to make sure no child

> is left behind, every child must learn to read. And every child must learn to add and subtract. (Applause.) So in return for federal dollars, we are asking states to design accountability systems to show parents and teachers whether or not children can read and write and add and subtract in grades three through eight.

(Bush 2002; emphasis added).

Seven years later, Arne Duncan, the Secretary of Education under President Obama, announced the launch of the federal Race to the Top (RTT) initiative. Building on the foundation laid by NCLB, RTT brings in some familiar-sounding elements. First, Duncan lays out a rationale for reform anchored in economic competitiveness:

> "We take our cue here from the president. He starts with the understanding that maintaining the status quo in our schools is unacceptable. He recognizes that America needs urgently to reduce its high dropout rates and elevate the quality of K-12 schooling—*not just to propel the economic recovery but also because students need stronger skills to compete with students in India and China.*"

(Duncan 2009, Emphasis added). Next, Duncan turned to accountability, the core lever of change:

> To boost the quality of teachers and principals, especially in high-poverty schools and hard-to-staff subjects, states and districts should be able to identify effective teachers and principals. At the local level we want to see better strategies in place to *reward and retain more top-notch teachers—and improve or replace ones who aren't up to the job*… [T]o turn around the lowest-performing schools, states and districts must be ready to institute far-reaching reforms, replace school staff, and change the school culture. We cannot continue to tinker in terrible schools where students fall further and further behind, year after year.

(Duncan 2009, Emphasis added). Finally, Duncan identified a key facilitator of change in the form of measurement and enhanced standardization:

> [W]e are looking for Race to the Top states to adopt common, internationally-benchmarked K-12 standards that truly prepare students for college and careers. To speed this process, the Race to the Top program is going to set aside $350 million to competitively fund the development of

> *rigorous, common state assessments*. Award-winning states will be able to monitor growth in student learning…

(Duncan 2009, Emphasis added).

Thus, like past reforms, current reforms are driven by fears about competitiveness, motivating calls for increased accountability, this time with an especially strong focus on teachers. As Taubman (2009) notes, current efforts derive legitimacy from their relationship to business and science. That was the case 100 years ago with the infatuation with scientific management. Similarly, although accountability efforts are focused to a high degree on teachers, as evidenced by the centrality of teacher evaluation system to most reforms, the history above shows attempts to do that dating back to the 1800s. The difference today is that statistical methods, data collection, and a more dominant instrumental rationality (see Rose et. al. 2006) have facilitated a more complete attempt to isolate individual impacts and thereby generate accountability through measurement.

1.1.3 Similar reforms, similar results

Given the similarities between the latest round of reforms and those that preceded it, it should not be surprising that research has failed to find evidence that these reforms are improving student outcomes at the hoped for rates. Two reviews punctuate this point, the Improving Teaching Effectiveness: Final Report on the Intensive Partnerships for Effective Teaching (Stecher, et. al. 2018) and Teacher Evaluation in American Schools, Rowan and Raudenbush's chapter in the 5th Edition of the Handbook of Research on Teaching (2016).[2] Both conclude that the impact on student outcomes of the most

[2] The Institute for Education Science's also financed an evaluation of Race to the Top overall. Completed in October 2016, the report found that while Race to the Top spurred

aggressive efforts to reform education through individual accountability have been underwhelming. Rowan and Raudenbush (2016) go further in explaining the challenges accountability schemes pose as a tool for systemic improvement in public education, though, as I argue in chapter 8, their approach may understate the limits of accountability.

Stecher et. al.'s (2018) report follows a six year evaluation of the efforts of three traditional public school districts in California and four Los Angeles-area CMOs. As part of the Intensive Partnership on Teacher Evaluation, all sites implemented new measures of teaching effectiveness and modified personnel policies by, for example, conditioning tenure on evaluation scores. All sites' evaluation systems also included observation rubrics to evaluate teacher practice and outcome measures based on standardized tests. Most observation rubrics were based on Charlotte Danielson's Framework for Teaching (FFT) with principals or other school administrators observing teachers between two and five times per year. For outcome measures, traditional public districts used value-added measures (VAMs) while the CMOs used student growth percentiles (SGPs). The results were disappointing. "…[M]easuring effectiveness and using it as the basis for teacher management and incentives did not appear to lead to gains in student achievement or graduation rates." (Stecher et. al. 2018, p. 498).

Acknowledging that their "evaluation does not tell us why these outcomes were not achieved," the authors were "willing to speculate—informed by our observations of

---

states to adopt most of the desired policies (adopting standards and assessments, reform of staffing practices for teachers and principals, improving conditions for charters and turning around low-performing schools) the impact on student outcomes was ambiguous. "Trends in student outcomes could be interpreted as providing evidence of a positive effect of RTT, a negative effect of RTT, or no effect of RTT." https://ies.ed.gov/ncee/pubs/20174001/pdf/20174001.pdf

the sites … during the past several years—about potential factors that might explain the lack of impact." (Stecher et. al. 2018, p. 498). Two hypotheses are particularly relevant here: the fact that summative measurement interferes with the ability to use evaluation information formatively and the resistance generated when trying to use evaluation scores to inform high-stakes decisions. As the authors put it, "it was difficult for the sites to navigate the underlying tension between using evaluation information for professional improvement and using it for high-stakes decisions. [In addition] some sites encountered unexpected resistance when they tried to use effectiveness scores for high-stakes personnel decisions… despite the fact that the main stakeholder groups had given their support to the initiative in general terms at the outset." (Stecher et. al. 2018, p. 500). Their final recommendation was to take a deeper look into the implementation of efforts like this to unpack the connection between reforms and outcomes: "In change efforts such as this, it is important to measure the extent to which each of the new policies and procedures is implemented in order to understand how the specific elements of the reform relate to outcomes." (Stecher et. al. 2018, p. 503).

In a more general multidisciplinary review of the use of teacher evaluation, Rowan and Raudenbush (2016) came to a similar conclusion. Finding that teacher evaluation has had disappointing results both in terms of improving the composition of the workforce and improving student outcomes, the authors dove deep into the personnel management literature to find an explanation. Relying on organization theory, principal-agent theory and organizational psychology, the authors argue that the problem is fundamentally one of measurability. Measuring teacher performance is difficult given the complexity and uncertainty of the technologies involved. As a result, measures of teacher

performance often fail to capture the totality of what we want teachers to do (formally, the measures are "distorted."). Likewise, there is a high risk of evaluations resulting in noisy or biased measures of even the limited portion of teachers' work they measure. Even the use of composite scores from multiple measures of effectiveness fails to avoid highly problematic probabilities of miscategorizing teachers on an ordinal scale. The authors thus conclude that the use of objective summative evaluation may be an inherently limited tool for reform. "When performance measures are risky and distorted, optimal personnel policy will down-weight the simple, formulaic use of 'objective' and 'quantifiable' performance measures in personnel decision-making and move instead to more 'subjective' performance appraisals that capitalize on supervisors' intimate local knowledge." (Rowan, B. & Raudenbush, S. 2016, p. 1161). Their reference to local knowledge here is almost in passing. The framework applied in this dissertation, however, suggests that local knowledge plays a critical role to the explanation of why education reform efforts have so often failed. This framework also suggests that Rowan and Raudenbush's conclusions about how problematic summative evaluation is may understate the degree of the problem.

1.1.4 The case for New Jersey cases

New Jersey is a compelling site for diving deeper into the underlying causes of this lack of success because, like all states, it adopted standardized tests following NCLB and, more recently, New Jersey aggressively pursued strategies similar to those reviewed in Improving Teaching Effectiveness and Teacher Evaluation in America's Schools. In fact, New Jersey uses one of the composite measures of teacher performance analyzed by Rowan and Raudenbush. As the authors noted, evaluation efforts that include teacher

observation and student outcome measures, as New Jersey does, had been tried in 40 states as of 2016. (Rowan, B. & Raudenbush, S. 2016, p. 1159).

A quick review of the development of New Jersey's evaluation system highlights the similarities between New Jersey and other states. With the passage of NCLB, New Jersey developed a new assessment for grades 3-8, the New Jersey Assessment of Skills and Knowledge (NJ ASK). New Jersey was then a regular applicant for RTT awards. While it struggled to win the awards, only doing so in a later round with lower stakes, it adopted several elements consistent with RTT's theory of change. In 2010, the NJ State Board of Education adopted the Common Core State Standards (CCSSs) that had been developed by the National Governor's Association and funded as part of RTT.[3] Around the same time, New Jersey joined the Partnership for Assessment of Readiness for College and Career (PARCC), a consortium of states developing an assessment aligned to the CCSSs. The resulting PARCC assessment replaced NJ ASK for ELA and Math in the 2014-15 school year. In 2012, two years after joining PARCC, New Jersey's legislature passed TEACHNJ, calling it tenure reform legislation. TEACHNJ required a new teacher evaluation system with four annual rating categories – Highly Effective, Effective, Partially Effective and Ineffective – based on multiple measures of student learning and instructional practice. AchieveNJ, the implementing regulation, defined the specific components of those measures for all districts across the state. To ensure implementation, AchieveNJ centralizes discretion over all districts' evaluation systems, subjecting them to annual DOE commissioner approval and exempting them from collective bargaining. It

[3] In 2016, the Christie administration announced revisions to about 230 of the over 1400 standards, naming the result the New Jersey Learning Standards.

also centralized the relative weights of the elements of the system. (Historical Context 2016)

AchieveNJ further established that educator effectiveness must be measured, that the measurement would include both student growth and instructional practice, that Student Growth Percentiles and Student Growth Objectives would measure student growth, and that instructional practice would be measured by three teacher observations according to a district-selected rubric. Principal practice would similarly be measured by observations and measures of student growth. While districts could select observation rubrics, those rubrics are subject to state approval. Likewise, while teachers create their SGOs subject to the approval of their supervisors, SGPs are based on state standardized tests and are designed and calculated by the state. (Historical Context 2016)

New Jersey is therefore a strong case for exploring the underlying causes for accountability-based reform's lack of success. Like the districts and CMOs in the Improving Teaching Effectiveness partnership, NJ districts, including the three that are the subjects of this dissertation, have adopted multiple measures of teaching effectiveness and formally linked those to HR decisions. New Jersey also has a wide spread between its top performing districts and its many struggling urban districts. As such, a framework that identifies reasons NJ's reforms might fail has a good chance of being useful to assess reforms currently in place elsewhere and those still to come.

1.2 Significance

The policy implications of the current slate of education reforms are difficult to overstate. The adoption and implementation of new curricula aligned to CCSSs, the development, distribution and grading of new standardized tests, the development and

execution of a complex, multi-measure teacher and principal evaluation system and the commensurate revisions to the tenure process, represent an enormous allocation of resources. As such, their success is critical to avoiding the waste of those resources. Conversely, the harm may not be limited to wasted resources if they fail to bring about substantial improvements as intended. Rather, the harm from unintended consequences of an effort to reengineer a process as social, complex and critical to both the millions of students and staff involved and to the broader social world that public education affects could set us back decades. For example, if reforms as currently constituted exacerbate rather than remedy historical gaps between different groups of students, we are wasting precious time needed to close the gaps for the current generation of students.

That the latest round of reforms has already shown limited effectiveness while seemingly repeating mistakes made over a century prior only adds to the urgency for better understanding why education reform so often fails to live up to expectations. Perhaps with a framework for understanding why reforms so often fail, future policymakers can avoid repeating the same mistakes in the inevitable next round of reforms.

1.3 Organization

In chapter 2 I develop the broader theoretical framework that I will test against three cases to learn whether it has the power to shed light on education reform's poor track record. Drawing on literature about modern efforts to govern complex social and physical processes, I argue that the education reforms studied here, as well as the history I summarized in section 1.1 are particular instances of general efforts to make the world legible to government. To the extent that this is true, we should be able to learn from the

consequences of past efforts that have been studied in detail and apply those lessons to the current reforms. This literature, bolstered by literature on education organizations and public service more generally, suggests a general explanation for the failure of past education reforms and offers a test to which we can subject the recent reforms to get a sense of why there has been so little evidence of success. The framework includes three conditions common to failed legibility efforts. It also identifies mechanisms by which those conditions lead to failure.

Chapter 3 recapitulates this literature review as the series of research questions I hope to answer in this dissertation. I also review the methods used to answer these questions, and introduce the three districts that are the sources of my local data.

I report my findings in chapters 4 and 5. In chapter 4, I apply the framework developed in chapter 2 to the documented details of New Jersey's education reforms and evidence from three New Jersey districts. I argue that these reforms satisfy the conditions for failure identified in chapter 2. In chapter 5, I review the evidence from all three cases to identify whether the mechanisms of failure are present. This includes the results of a quantitative analysis aimed at confirming whether one particularly measurable consequence, teacher turnover, is being impacted as might be expected. In chapter 6, I review evidence from the three districts to speculate about whether any broader unintended consequences might be likely. I then conclude by revisiting my findings from a different angle, asking "is there anything we can learn from this and past instances that might suggest a better way forward?" Answering in the affirmative, I propose guidelines for the next round of reform, arguing that avoiding the three conditions of failure might go a long way to breaking the cycle of failed reform efforts.

2 Lessons from Failed Efforts to Simplify the World

The characteristics of recent education reforms are consistent with historical efforts to reform education. Insofar as they aim to govern education by measuring outcomes, these reforms are also consistent with failed modern efforts to govern the social and physical world in other fields. Following Scott (1998), I refer to such efforts as "legibility efforts" throughout this dissertation because they are fundamentally about making the complex and remote "legible" to government. Patterns in other failed legibility efforts suggest three conditions are always present. I propose these conditions as a framework for better understanding the persistent ineffectiveness of accountability-driven education forms. Combining this framework with past research into the nature of public service and public education leads to straightforward questions: (1) does education reform satisfy all three conditions of failure and, if so, (2) by what mechanisms might these conditions be leading to the failure of reforms?

2.1 Other efforts to make the complex governable demonstrate the conditions of failure

Based on the idea that understanding what governments do and how they do it is more important – and more possible – than theories of state action in the abstract, Foucault's governmentality is concerned with the ways in which the social and natural world are made governable for those who aim to govern and for the governed themselves. (Foucault 1991). Initially a genealogical study of changes in the practice of government, Foucault's ideas have formed the basis for an analytics of government by those who followed him, a lens that analyzes government action through the technologies used to effectuate control over a population. (E.g. Rose and Miller 1992).

As explicitly stated by the Bush and Obama administrations, the core of recent education reform efforts is a means of measuring performance – of students, teachers and principals primarily - and a means of holding individuals – teachers and principals primarily - accountable for that performance. Implicit in these reforms is the need to generate knowledge about what is happening in schools and classrooms. Creating this knowledge requires tools that will render the complexities of education knowable by governmental actors. A lens that analyzes government action through the technologies used to control the governed is therefore particularly useful to the study of both past and current education reforms.

The technologies used to create legibility in other fields - categorization, sorting, standardization and calculation – have been the subject of many of the studies building on Foucault's ideas. (Rose 1988, Hacking 1991, Rose 1991, Scott 1998, Bowker & Starr 1999, Mitchell 2002, Rose et. al. 2006, Miller and Rose 2008, Dean 2010). Identifying the characteristics that led to the failure of past legibility, these studies offer a particularly useful frame to guide the analysis of the technologies of education reform. There are three common conditions of failure across these studies: (1) simplification of a complex social or natural process, (2) a shift of expertise from local control and situated knowledge to centralized experts and (3) ignorance of context.

Three studies are particularly instructive in highlighting these three conditions. In "Sorting Things Out," Bowker and Starr (1999) analyze the extreme legibility effort required to effectuate apartheid in South Africa. In "Seeing Like a State," Scott (1998) reviews several cases, from forestry in 18th-century Prussia and Saxony to compulsory

villages in Tanzania. In "Rule of Experts," Mitchell (2002) analyzes several cases from 20th century Egypt. Scott (1998) might best sum up the theme of these studies:

> The metaphorical value of this brief account of scientific production forestry is that it illustrates the dangers of dismembering an exceptionally complex and poorly understood set of relations and processes in order to isolate a single element of instrumental value… Utilitarian simplification in the forest was an effective way of maximizing wood production in the short and intermediate term. Ultimately, however, its emphasis on yield and paper profits, its relatively short time horizon, and, above all, the vast array of consequences it had resolutely bracketed came back to haunt it.

(Scott 1998, p.21).

2.1.1 Failed legibility efforts oversimplify complex processes

Those studying governance through the lens of governmentality argue that governing large populations, nature or complex processes requires acts of simplification. "No administrative system is capable of representing any existing social community except through a heroic and greatly schematized process of abstraction and simplification." (Scott 1998, p. 22). This is a specific instance a much broader truth: simplification is a necessary feature of all science and all policy as each requires simplifying models to make them comprehensible and actionable, or, in Scott's words, legible. The utility of modeling and thus simplification is significant. Simplifying models made possible welfare-increasing technologies from water filters to GPS and policies many would consider highly successful such as the Clean Water Act. The argument, then, is not that all simplification leads to failure. Instead the issue is with the degree and/or nature of simplification. The degree of simplification depends on the complexity of the thing being simplified and how complex the model can be to remain useful. The nature of simplification depends on what is being simplified and how that simplification is effectuated.

Because all science and policy requires simplification, increasing complexity poses increasing challenges. This is reflected in what is sometimes referred to as Bonini's paradox. Accurately modeling something complex requires a complex model, but more complex models are generally less useful (by being, for example, less understandable). Thus, more accurate models are often less useful and, as a corollary, the more complex the process being modeled, the greater the gap between utility and accuracy. (Bonini, C.P. 1963). From a policy perspective, the implication is that attempts to simplify more complex aspects of reality entail more risk. If a policy requiring a high level simplification is applied to a highly complex natural or social process, the gap between the model and reality will be large.

The nature of simplification, on the other hand, depends less on complexity than on the object being modeled. Here the legacy of Foucault is particularly instructive. Because simplification often takes the form of counting and calculation, it generally includes governmental technologies necessary to make something countable and calculable. (Rose 1988, Hacking 1991, Rose 1991, Rose and Miller 1992, Rose et. al. 2006, Scott 1998, Mitchell 2002, Miller and Rose 2008). To make things countable in turn requires the creation of categories; once categories are created, things and people can be counted according to their membership in a particular group. (Hacking 1968, 1991, Porter 1995, Bowker & Starr 1999). Counting the number of people who are insane, for example, requires the creation of categories of sanity. People, however, are imperfect objects of science and policy because they are also subjects; they respond to and are affected by being categorized in unpredictable, often problematic, ways. (Rose 1988).

The implication for policy is that human behavior poses unique risks and challenges. As with complexity, the more simplification required, the greater the risk.

Several examples illustrate the point that simplification becomes problematic as a function of the degree or nature of simplification. In their analysis of apartheid South Africa, Bowker and Starr (1999) provide an example of a simplification that was problematic because it oversimplified something highly complex, because the objects of simplification were people and because the means by which that simplification was effectuated required arbitrary categories. The authors note in particular the role categorization played in enforcing apartheid's rules of separation. To assign different rights and privileges to different groups, administrators needed to be able to assign individuals to the groups. Doing so further required having clearly defined groups. Both tasks proved exceedingly difficult. Two examples highlight the extremity of the need for simplification. For administrative purposes, the easiest means to distinguish groups was by physical characteristics. Two of the characteristics were skin color and hair "kinkiness." To test skin color, officials created a color chart ranging from lighter to darker colors. Bureaucrats were tasked with comparing individuals' skin to the colors on the chart. Their racial classification depended in part on the color the bureaucrat identified. To test hair kinkiness, bureaucrats sometimes resorted to the pencil test. In this test, an individual would put a pencil in the hair and then face downward. If the pencil fell out, the person was classified as white. If not, they were classified as colored or black. As a result of these classification instruments, it was not uncommon for family members with different skin tones or hair textures to be classified into different groups and have different rights. As the author's note, "[f]or a bureaucracy to establish a smooth

data collection effort, a means must be found to detour around… higher order issues." (p. 24).

Scott (1998) identifies legibility, the taking of complex, local social practices and creating a means of central recording and monitoring, as the central problem of statecraft. He notes that all such efforts entail simplifications that do not - and are not intended to - match reality but instead represent only the slice of reality that is of interest to the observer. While he uses several examples, he begins with scientific forest management as the model that provides the lens through which he analyzes other efforts from urban planning to rural settlement and agriculture. Scott highlights the key driver of simplification in scientific forestry: the early modern European state viewed forests in terms of "the revenue yield of timber that might be extracted annually." (Scott 1998, p.12). To maximize this number, the state reduced the complexity of the forest ecosystem into a single value: the volume of lumber or firewood the trees could generate. The process of calculating the salable volume of wood involved further simplifying assumptions about the trees. The result was a highly simplified forest monoculture. Unfortunately, "a whole world lying 'outside the brackets' returned to haunt this technical vision."(Scott 1998, p.20). This example highlights the degree of simplification more so than the nature of it; the forest ecosystem is highly complex and its managers aimed to control it with a single number.

Finally, Mitchell (2002) offers several examples of simplification. I focus on one such example here: the creation of a map of Egypt. The focus on maps is important as "the map signifies the massive production of knowledge, the accuracy of calculation, the entire politics based upon a knowledge of population and territory that Foucault

characterizes as governmentality…" (Mitchell 2002, p.18). After Great Britain invaded Egypt in 1882, one of its first priorities was "a vast project of calculation." (Ibid.). To raise revenue, Britain needed to "determine, for every square meter of the country's agricultural land, the owner, the cultivators, the quality of the soil and the proper rate of tax." (Ibid.) To do this, they developed the first comprehensive land map of Egypt. "The map was intended not just as an instrument of administrative control… but as a means of recording complex statistical information in a centralized…and visual form." (Ibid.) While the map was a celebrated technical feat, allowing the central authority a means to oversee the entire country, it was, by necessity, highly simplified. In particular, it lacked any non-spatial information and ignored all social, agricultural and other practices that were contained in those spaces. In short, "the map did not produce a more accurate or detailed knowledge of its object than earlier forms of governmental practice." (Ibid.)

These three examples highlight an additional aspect of simplification that merits a brief mention. While the initial discussion of the utility of simplification in science and policy hints at a benign, technocratic rationale, that does not mean simplification is free from politics or power dynamics. Power dynamics are, in fact, central to Mitchell's (2002) argument. A deep treatment of the subject is beyond the scope of this dissertation. As noted, my focus is on how the results of policy decisions are experienced by the actors that are the objects of reform. However, the two are inseparable in at least one respect. I argue that failure is a function not of simplification per se but of the degree and nature of simplification. These two characteristics of simplification are both a function of power dynamics and provide opportunities for power-dynamics to self-perpetuate. They are a function of power dynamics insofar as the agenda driving the policy that necessitates the

simplification is politically set. The need to categorize people in apartheid South Africa is an egregious example, as such categorization would not have been necessary but for the existence of the extremely political policy of apartheid. Similarly, even otherwise benignly motivated simplifying acts open up the space for power dynamics to intervene. Once the decision to simplify is made, the choices about how and how much to simplify - what gets counted, what does not – and how to measure the outcomes are highly political. Moreover, the facial objectivity of simplifying acts can mask political choices and the underlying motivations, further enabling them. (See, e.g. Flyvberg 1998). I turn to some possible implications of this in the context of education reform in chapter 6.

2.1.2 Failed legibility efforts shift expertise from locals with situated knowledge to centralized experts

The second common condition of failure is the shift of expertise from local control and situated knowledge to centralized experts. The three examples from the prior section are again instructive. The Bowker and Starr (1999) example is perhaps the most extreme. The shift of expertise about one's heritage was exemplified by the occasional use of barbers to analyze individuals' hair for purposes of determining their classification. The simplification of classifying an individual by their hair led to a shift in expertise from the highly local (an individual's own knowledge of their ancestry) to barbers who were "expert" in hair. Perhaps more analogous to present acts of simplification, the definition of groups was centralized, with expertise residing entirely with the bureaucracy that created both the group definitions and the means of testing needed to assign individuals to those groups.

Scott's (1998) example of 18th century scientific forestry further highlights the tendency of legibility efforts to centralize expertise at the expense of local, situated knowledge. "At the limit, the forest itself would not even have to be seen; it could be 'read' accurately from the tables and maps in the forester's office." (p.15) Yet the ability to manage forests remotely was but one of the expertise-shifting impacts of scientific forestry. Another more directly impacted the expertise required to work in the industry by making a highly complex process routine:

> With stands of same-age trees arranged in linear alleys, clearing the underbrush, felling, extraction, and new planting became a far more routine process. Increasing order in the forest made it possible for forest workers to use written training protocols that could be widely applied. A relatively unskilled and inexperienced labor crew could adequately carry out its tasks by following a few standard rules in the new forest environment.

(Scott 1998, p.18). Thus a complex process dependent on local, situated knowledge, could be carried out by a few central bureaucrats guiding an unskilled labor force.

The centralization of expertise might be most stark in Mitchell's (2002) map-making example. This is not merely because of the knowledge-centralizing effect of a small physical depiction of an otherwise illegibly massive space. It is also because of who created the map, for it was their expertise that dominated. Local surveyors of each plot of land did not create the map, imbuing it with their situated knowledge. Instead, the map was the work product of the Ministry of Finance, who hired a British captain and British assistants to lead the survey work. The maps launched the "continuous and systematic government production of statistical knowledge" in Egypt. (Mitchell 2002, p.77). But while it launched the era of government-generated statistics, it was not the first modern map in Egypt nor was it the first effort to calculate land area for

administrative purposes; ancient Egypt had been doing so for centuries. Instead, it was the way it was created, and the expertise that was required, that distinguished this effort. The map was based on triangulation, a technique that required specific technical expertise to subdivide the entire country into small, equally sized sections. A British geography journal celebrated that the map "would be based for the first time on a 'rigorous framework.'" (Mitchell 2002, p.79). This idea of rigor being defined by outside expertise and the rendering of the complex calculable is particularly resonant in present discourse surrounding education policy.

In Egypt, this "rigor" replaced the local knowledge contained in the old cadasters, and with it, fundamentally changed the relationship between government and the governed:

> "The old cadastre was assembled from a knowledge of households and villages. Land claims and tax liabilities were the claims and liabilities of communities of persons, and expressed the relations of those communities both to the land and to those in power. Movements of information, revenue, and control flowed through these relations. Under the new system, the list of persons was merely 'complementary' to the map, supplying additional information 'that could not conveniently be inserted on the plan of a piece of land.'

(Mitchell 2002, p.80). In addition to displacing local knowledge, the map "moved the site where all this knowledge was held." (Mitchell 2002, p.81). Rather than being held by village surveyors who had a "vital skill," detailed knowledge of how the Nile's floods impacted the local land, the map was held by central bureaucrats.

Centralization of expertise is also a historical theme in education. For example, Grant's (1989) characterization of past accountability-driven reforms in education echoes the flags raised by Scott and Mitchell. Noting that when a "model reduces felt reality to a

series of abstractions...[it is] vital to ask how this reduction is accomplished." (Grant 1989, p.156), Grant argues that the "central tension in the process of one group modeling the activity of another is grounded in different views of the nature of the education process, the weight of teachers' experience and practical wisdom versus evaluation models." (Grant 1989, p.157). Rowan and Raudenbush alluded to this tension in their recommendation that performance measures leverage "supervisors' intimate local knowledge." (Rowan, R. and Raudenbush, P. 2016, p. 1161).

As with oversimplification, the shift of expertise is not free of politics or power dynamics. In fact, this shift in expertise is one of the ways in which simplification opens up the space for power dynamics to intervene under cover of technocratic decision-making. In the case of scientific forestry, for example, the shift of expertise to a central administrator and the creation of conditions under which a less skilled workforce can harvest wood can appear to be rational byproducts of an otherwise objective policy goal. It is not hard to imagine, however, the deskilling of the workforce being a welcomed benefit for powerful stakeholders. While this analogy extends naturally to education reform – one can imagine stakeholders who might welcome the devaluation of teachers as professionals – I do not address these questions in this dissertation beyond consideration of whether and how teachers experience devaluation of their expertise and some potential implications of that.

2.1.3 Ignorance of context

The third condition of failed legibility efforts is ignorance of context. As a practical matter, ignorance of context is highly related to oversimplification; cases of oversimplification will most often involve policymakers ignoring context. However, I

treat them separately here because oversimplification and ignorance of context capture different aspects of what legibility efforts ignore. Oversimplification addresses outcomes and more specifically the degree to which policymakers select a narrow outcome and a narrow way of measuring it from a complex set of possible outcomes and measures. Ignorance of context addresses not outcomes or even possible outcomes but how those outcomes and measures are impacted by the setting in which they exist. This distinction has utility because a focus on context is not restricted to the framing of the problem; in fact it is explicitly a reminder to broaden the frame as much as possible. In evaluating policy choices, it opens up the space to consider what, besides outcomes, policymakers might have ignored and why. In designing policy, it serves as a reminder to consider not just what outcomes matter, but what conditions those outcomes might depend upon. As the next example highlights, those conditions are often policy choices themselves and could have substantial implications for a wide range of outcomes.

Mitchell's (2002) description of the international response to Egypt's 20th century "food shortage" provides a good illustration of the utility of distinguishing ignorance of context from oversimplification. In an effort to address food shortages in Egypt in the late 20th century, international institutions focused, in part, on population growth, arguing that food was short because the population was growing faster than Egypt's Nile-fed arable land could support. Through this lens, Egypt's population growth rate of 2.5% was deemed too fast. Within this narrow frame, the rural poors' average of 7.5 children .seemed to validate "expert" concerns about the ability of the land to sustain the population. As Mitchell notes, this oversimplification ignored critical aspects of the social context of Egypt at the time, many of which were theoretically malleable policy

choices rather than fixed features of geography. In particular, the narrow framing ignored how a lack of social security, the male dominated economy and high childhood death rates drove population growth. The lack of a social security system and the fact that only male children were likely to earn incomes meant that having two surviving male children might be the best way to ensure parents were supported in old age or illness. At the same time, around 1 in 3 children died in childhood. In this context, having 7 or 8 children is not excessive; it's rational insurance. This highlights the utility of looking at context independent of outcome selection. Consideration of context opens up a totally different and potentially more just set of policy options. Rather than focusing on food production and trying to limit birth rates amongst the poor, might policymakers have tackled social security, youth mortality or the male dominated social order to greater social benefit?[4]

Within the education space, Sizer (1984) was particularly concerned with accountability-driven reform's tendency to overlook "special local conditions." To be sure, the education context would seem to pose a serious threat to reform efforts, especially given evidence of failure from decades of past reforms. As Lortie (1969) argued, contextual differences make education hard to rationalize. Building on Thompson's (1967) argument that evaluation practices are context dependent,[5] Lipsky (1980) laid out conditions for a successful accountability policy for public organizations.

---

[4] That these are far more politically difficult to tackle is not lost on me. This at least partially illustrates one reason oversimplifying legibility efforts are so appealing to policymakers and even social scientists. While a full consideration of these political dynamics is beyond the scope of this dissertation, I will touch on these topics in more detail in the conclusion.

[5] More specifically, Thompson (1967) argued that evaluation practices depend on the degree to which goals can be clearly defined and the extent to which the organization has control over the relationship between inputs and outcomes. Lipsky's (1980) conditions incorporate these two.

According to Lipsky, successful accountability schemes require clear goals, performance that can be measured validly, and limited variability in context. Rowan and Raudenbush (2016) essentially hit on all these points, showing that the goals of education are too complex to define clearly, performance is extremely difficult to measure validly and context and student differences create issues for even the most complex instruments.

2.2 How might the three conditions of legibility failure explain the direct causes of education reform's disappointing results?

Literature from other fields, public service in general and education research in particular provide insight into the means by which these three conditions might lead to failure. Bowker and Star (1999), Scott (1998) and Mitchell (2002) provide general guides. But the patterns they identify show up repeatedly in modern efforts to reform public service delivery. Lipsky (1980) generalizes these for "street level bureaucrats" (SLBs) while the education literature referenced in the introduction along with examples from Ingersoll (2003) highlight implications for teachers as particular SLBs.

A key premise of this chapter is that the three examples relied on so far all resulted in failure. Insofar as Bowker and Star (1999) described the administrative apparatus necessary to sustain a patently unjust policy that no longer exists, its failure should be obvious. The failures from Scott (1998) and Mitchell (2002) are subtler. Scott (1998) found that scientific forestry was a failure after first being a "resounding success." (p.19). "It took about one century for them [the negative consequences] to show up clearly. Many of the pure stands grew excellently in the first generation but already showed an amazing retrogression in the second generation." (Scott 1998, p.22). This retrogression included production losses of up to 30%. A key here was that the nutrient

cycle – consisting of complex relationships between "soil building, nutrient uptake, and symbiotic relations among fungi, insects, mammals, and flora" was disrupted by the clearing of what was seen as extraneous material while the monoculture made the trees highly sensitive to blight. "A new term, Waldsterben (forest death), entered the German vocabulary to describe the worst cases." (Scott 1998, p. 20)

He suggests with the power of hindsight, however, that failure might have been foreseeable if administrators had a proper appreciation for the complexity of their undertaking:

> "This utopian dream of scientific forestry… was not and could not ever be realized in practice. Both nature and the human factor intervened. The existing topography of the landscape and the vagaries of fire, storms, blights, climatic changes, insect populations, and disease conspired to thwart foresters and to shape the actual forest. Also, given the insurmountable difficulties of policing large forests, people living nearby typically continued to graze animals, poach firewood and kindling, make charcoal, and use the forest in other ways that prevented the foresters' management plan from being fully realized."

(p.19).

In the example of map-making from Mitchell (2002), the exercise failed in that "the calculations that it was supposed to enable were never quite made possible." (p.18). "Lyons could not claim, in fact, that the survey he directed resulted in a more accurate measure of the land… the new maps were in significant ways less accurate and more cumbersome than the old methods of recording landholdings." (Mitchell 2002, p.82). Not unlike a short-term look at the success of scientific forestry, however, the failures of the new map were obscured if one did a traditional evaluation of the map. "Thanks to the gap opened up between field and map, the question of accuracy could now be recast. It was now an issue of one, simple relationship: the correspondence between the map and 'the real world.'" (Mitchell 2002, p.82). This suggests that using the measurement instruments

created to effectuate accountability policy to evaluate that policy might obscure failures or risks. The framework developed here avoids dependency on oversimplifying metrics, which in any event can only identify failure, not explain it.

From the literature, I derive four categories of mechanisms relevant to education reform. First, simplification, centralization of expertise and ignorance of context often lead to a mischaracterization of the problem. It is almost axiomatic that solutions misaligned to the real problem have little chance of success, and the evidence bears this out. Second, the three conditions lead to resistance by the populations they are trying to govern – as Stecher et. al. (2018) found - or at least trigger unproductive coping mechanisms. This resistance is rarely one of the behaviors the theory of action seeks to bring about. Third, by devaluing situated knowledge and ignoring context, legibility efforts tend to undermine other conditions of success. Finally, oversimplification, ignorance of context and shifting expertise can lead to what administrators will recognize as poor implementation.

2.2.1 Mischaracterization of the problem

One necessary condition of the success of any solution is that it addresses the actual cause of the problem. Yet proper diagnosis of the problem is challenging where the underlying issue has been highly simplified, excludes local knowledge and ignores context. As such, failure by virtue of mischaracterizing the problem is perhaps the most foreseeable source of failure.

The food shortage case from Mitchell (2002) highlights this well. Mitchell details how simplifying efforts driven by international experts led to a serious mischaracterization of the problem. Conventional wisdom characterized the problem as a

simple one whereby the narrow strip of arable land around the Nile was insufficient to support a fast growing population of millions of people. In this characterization, there were two problems: a lack of fertile land and population growth. However, data showed that agricultural productivity outpaced population growth during the time in question. "In 1991, food production per capita was 17 percent higher than at the start of the previous decade. So it is not true that the population was growing faster than the country's ability to feed itself." (Mitchell 2002, p.172). The problem, instead, was very different, and implicated complex power relationships. Rather than there not being enough food, it was the distribution of food, and the demand for more resource intensive foods from wealthier Egyptians, that drove the shortage: "It was the switch to meat consumption, rather than the increase in population, that required the dramatic increase in imports of food, particularly grains." (Mitchell 2002, p.173).

To help frame the analysis of education policy, its worth going beyond the mischaracterization to understand the source of the mischaracterization and why, despite being relatively simple to dispel with data, it was so well accepted by experts and policymakers. As Mitchell notes,

> Open almost any study of Egypt produced by an American or international development agency and you are likely to find it starting with the same simple image. The question of Egypt's economic development is almost invariably introduced as a problem of geography versus demography, pictured by describing the narrow valley of the Nile River, surrounded by desert, crowded with rapidly multiplying millions of inhabitants.

(Mitchell 2002, p.176). One reason may be utility; a simple, visually appealing explanation allows policymakers to turn to solutions more quickly and without addressing potentially challenging details. In Egypt, the simple narrative allowed policymakers to avoid the sticky issue of land access and equity: "The image of a narrow

strip of fertile land crammed with so many millions of inhabitants enabled most contemporary analyses of Egyptian economic development to move very quickly past the problem of access to land. With so many people occupying so little space, the problem appeared to be already explained." (Mitchell 2002, p.176). Simple explanations that match the desired solutions of those in power serve a valuable rationalizing purpose. (See Flyvberg 1998).

However, Mitchell suggests that the dominance of the simple narrative may be part of a broader, methodological issue with social science. Simplification is helpful not just to policymakers in power but also to analysts. He notes that it is common for fields to rely on conventions for introducing problems. These conventions often become "tropes [that] seem too obvious and straightforward to question." (Mitchell 2002, p.169). Tropes predominate because of how much easier they make it to set up an analysis. According to Mitchell, the imagery launching analyses of Egypt does this in two ways, both serving to form it as an analyzable object. First, "the topographic image of the river, the desert surrounding it, and the population jammed within its banks defines the object to be analyzed in terms of the tangible limits of nature, physical space, and human reproduction." (Mitchell 2002, p.170). Second, the "naturalness" of the simple image sets it up as an object separate from and external to the study. He suggest that this serves the interests of international development experts who have "a special need to overlook…internal involvement in the places and problems [they] analyze, and present [themselves] as an external intelligence…" (Mitchell 2002, p.176). Tropes help facilitate this.

The narrative in the introduction suggests that education reform may be connected to the same type of simplifying trope. Taubman (2009) refers to this in her treatment of how a test-supported audit culture came to dominate education policy discourse. One of the more dominant narratives in education is that the education system is failing, teachers are the most important contributors to student learning and therefore teachers are the cause of failure.[6] This creates as the object of analysis the reasons for teachers' failure. Given the focus on accountability, the implied characterization of the problem is that teachers fail because they lack the incentive to succeed. However, as Ingersoll (2003) notes, this is a highly questionable premise, calling "the unusual degree of commitment of those who enter the profession" a "valuable resource" that accountability policies might squander. (p. 236). In fact, rather than having insufficient motivation, teachers may tend to subject themselves to demands that are higher than are achievable. (Lortie 1975). Ultimately, this narrative and the corresponding test-driven accountability systems "abstract from the impossibly complex world of schools and education a virtual world, often represented in charts…, in which subjects [are] rendered visible, calculable, self-regulating, governable…and…commodifiable." (Taubman 2009, p.95).

2.2.2 Engendering resistance

Mischaracterization of the problem is only one mechanism by which legibility efforts might fail. Another is that by ignoring local expertise and the human context in which policies play out, failed legibility efforts engender resistance amongst affected populations. From our examples so far, Scott (1998) offers the clearest illustration of this. The simplification of the forest to maximize production of one item ignored the many

[6] The more complete narrative includes school characteristics, student characteristics, family and neighborhood contexts, broad economic factors and myriad other factors.

uses local communities had for the forest beyond wood harvesting. As a result, it was partially brought down by an inability to manage what might have been a predictable response from those local communities. Dependent on the forest for food and energy, they continued using it to graze animals and cut firewood.

The public service and education literature offer insight into how legibility efforts, especially those directed at accountability, lead to the type of resistance Stecher et. al. (2018) found. While resistance can manifest itself in several ways, for this dissertation I address three. One is intentionally not complying with the formal demands of the central authority. Loose-coupling is an example of how powerful this form of resistance can be. A subtler version of resistance is gaming or cheating – taking advantage of the difficulty of creating perfect measurement instruments to comply with the letter of the law while violating its spirit. Where key actors are employees, they can also resist by resigning. Finally, while technically not resistance in that it is not intentional, staff may inadvertently resist by resorting to unproductive coping mechanisms (or "satisficing"). The literature offers at least two mechanisms by which simplification, devaluation of local expertise and ignorance of context inspire resistance. First, key actors may resist when they do not perceive the technologies through which policy changes are effectuated as valid. Second, accountability policies might engender resistance when they threaten the value proposition of the profession.

Loose-coupling is perhaps the most dominant form of resistance in public education. Originally defined by Weick (1976) and further developed by Meyer and Rowan (1978, 2006), loose coupling offers a powerful, if partial, explanation for the overall stability of educational practice and results in the face of frequent, sometimes

aggressive, efforts to change it. In fact Weick (1976) opens with the same Stephens (1967) quote used in the introduction to this dissertation. As Weick (1976) describes it, entities are loosely coupled when they are responsive to each other but preserve their respective identities. The attachment between loosely-coupled entities is circumscribed and weak, and each is often slow to respond to changes in the other. For example, schools may be only loosely coupled to district administration, or teachers to principals. This serves several possible functions according to Weick. For example, loose-coupling supports stability by preventing organizations from having to respond to every whim of leaders in a field where leaders change often. It also allows local adaptation even in the face of formal standardization. Each of these functions of loose-coupling is consistent with treating it as a form of resistance; the product of efforts made by individual actors to avoid change in the face of external demands to do so.

Another form of – unintentional – resistance is satisficing, resorting to coping mechanisms to "get by" rather than fulfilling the spirit of bureaucratic requirements. This is often a symptom of a policy designed without regard to local knowledge and inconsistent with the context in which that policy will play out. Lipsky (1980) identifies several triggers of coping mechanisms. One trigger is stress and anxiety. Being evaluated generally adds stress. Because of the involuntary nature of their clients, when SLBs are evaluated by client performance, it amplifies that stress. The pressure to establish deference and obtain client compliance may thus trigger coping mechanisms. Burdensome housekeeping is another trigger. Accountability policies that impose significant housekeeping burdens by, for example, adding a great deal of paperwork, can

further constrain workers and trigger the use of coping mechanisms. One example is simply filling out all forms the same way to reduce the time and energy spent on them.

Ignorance of context may play a huge part in the extent to which school staff resort to coping mechanisms. While researchers are concerned with the burdens – record keeping and otherwise - on teachers and staff in all schools, those burdens might be different in different contexts. As with resistance, for example, we might expect to see greater reliance on coping mechanisms for teachers in higher-poverty schools.

The additional burden on principals may make this worse. Because principals are the primary observers of teachers, teachers' perceptions of the quality of observations play a role in teachers' response to being evaluated. Unfortunately, principals and vice-principals are not equally competent to conduct observations of teachers, calling into question the validity – and especially the comparability across schools - of observation ratings. This was a key factor in Rowan and Raudenbush's (2016) analysis. Time is likely to play an amplifying factor in this: not all principals and vice-principals have the same number of teacher observations to conduct. Moreover, many Teacher Evaluation Systems require more observations of inexperienced or low-performing teachers,[7] which means principals with more struggling teachers will have more observations to conduct. Because inexperienced and low-performing teachers are unequally distributed in high-poverty, high-minority schools (Guarino et. al. 2006, Johnson et. al. 2005, Jacob 2007, Borman et. al. 2008), principals and vice-principals in those schools will be subject to higher burdens. Moreover, because these schools are more likely to be underperforming and suffer from organizational and resource deficiencies, both the stress and the difficulty of

[7] New Jersey is one of these.
http://www.state.nj.us/education/AchieveNJ/teacher/overview.shtml

managing the extra burden will be higher than in higher-performing schools. It is therefore reasonable to ask whether principals in low-performing schools will provide the same quality of evaluation and feedback as principals in high-performing schools. They may simply develop satisficing routines whereby they rank all teachers similarly average to save time, a coping mechanism that would manifest in rating "compression." Alternatively, principals may develop satisficing routines whereby they divert attention from other tasks to accommodate their increased commitment to observations, and management may suffer.

Gaming is a more acute form of resistance. The most visible version of this is "teaching to the test" but Atlanta's cheating scandal offers a more extreme version. Equally extreme but less nefarious, turnover is in a way the ultimate form of resistance. Turnover is a real problem with real costs. (Lipsky 1980, Ingersoll 2003). In education this goes beyond the general proposition that job satisfaction is related to turnover. (Perrow 1986). Teaching has been subject to chronically high turnover for decades. (Lortie 1975, 2003). It is, however, generally lower where teachers have more control. (Ingersoll 2003).

That teachers are less likely to leave when they have more control hints at one reason accountability efforts might trigger resistance. In his comprehensive analysis of public service, Lipsky (1980) defined SLBs as public employees that interact directly with citizens and have substantial discretion. Lipsky and others (e.g. Wilson 2000) have highlighted that attempts to manipulate the behavior of SLBs from above through accountability measures are fraught with challenges and the results are far from guaranteed to match the intentions. The engendering of resistance is one reason for this.

Because discretion plays a critical role in promoting self-regard and legitimacy, and because self-regard and legitimacy are key elements of the value proposition of public service, an accountability policy that either directly limits discretion or alters organizational culture in a way that affects workers self-perception can engender resistance. (Lipsky 1980, Wilson 1989).

Intrinsic rewards are not affected by public esteem alone, however. Another source of esteem comes from the discretion and autonomy many argue is critical to teaching. The desire to retain autonomy, which is not institutionalized, has been advanced as an explanation both for the loosely coupled structure of education organizations discussed above, and for concrete positions such as teachers' resistance to merit pay and standardized testing. Teachers "can make the most of transitive rewards only if there's freedom for them to choose the criteria and technology to be used in assessing student performance." The flow of intrinsic rewards is contingent on self-perceived classroom achievement, thus depending on personal goals, and affects other aspects of occupational life. (Lortie 1969). While the amount of autonomy teachers actually have is a debatable empirical question, there are many aspects of teachers' jobs that are outside their control. (Sizer 1984, Ingersoll 2003). The curriculum is often given, time is structured by the administration, teaching materials are given, school structure and administrative structure are set elsewhere and teachers do not really choose colleagues. Taking more aspects of the job from teacher's control might threaten the value of the job enough to trigger resistance.

Ingersoll and Lipsky also suggest that perceptions of validity of the instruments used to create accountability can amplify resistance. Validity is a particular threat where

control is highly centralized, as it likely devalues local expertise and ignore local context. This makes it likely to, for example, hold teachers accountable for things they don't actually control. (Ingersoll 2003). Again, this may be underlying the findings of Rowan and Raudenbush (2016) and Stecher et. al. (2018).

It is worth noting here that we might not expect resistance to look the same in all places. A key part of the context of education reform is that not all schools, districts, teachers and principals are created equal. Because teachers do not face identical challenges, imposing identical requirements actually imposes higher burdens on those teachers who need the most help. Because inexperienced and low-performing teachers are unequally distributed in high-poverty, high-minority schools (Guarino et. al. 2006, Johnson et. al. 2005, Jacob 2007, Borman et. al. 2008), teachers in those schools are likely to be more burdened than teachers in better-off schools. We therefore might expect to see more evidence of resistance in higher-poverty schools. For example, given the relationship of working conditions to turnover noted by Ingersoll (2003) and the difficulties of staffing high-poverty schools, we might expect to see greater turnover in those schools. On the other hand, we might expect to see loose-coupling preserved better in districts less dependent on external funding and with higher-performing students.

2.2.3 Undermining of the Conditions of Success

Legibility efforts that oversimplify complex processes, devalue local expertise and ignore context may also undermine other conditions critical to success. Given that the conditions of success are so often the subject of local knowledge and context, modern efforts that aim to create legibility through calculation create a high risk that these will be ignored and potentially become casualties of misguided policies. When it comes to public

employees, and teachers in particular, there are several conditions of success that might be overlooked and impacted by simplifying reforms: legitimacy, motivation, control and flexibility and collaboration. As Ingersoll (2003) notes, “too much centralized control of teachers' work may undermine good teaching and demotivate, antagonize and ultimately drive out teachers." (p.218).

The Relationship Between Legitimacy and Authority

Lipsky identified several conditions as complicating the work of SLBs. These are amplified for teachers. Like other SLBs, teachers serve two clients: the students and the public. The primary clients of all SLBs are generally involuntary; they don’t generally choose to interact with the bureaucracy in the way they choose which restaurant to go to. However, in addition to being non-voluntary, teachers' student clients are also highly immature. The immature student is the key worker in the school; it is ultimately up to them to demonstrate that the goals of education have been met (i.e. that they have learned something). Eliciting cooperation is therefore critical and particularly challenging. (Ingersoll 2003). Moreover, cooperation is necessary both for classroom management and for instruction. Bidwell (1965) calls this the teachers' dual role dilemma: they are responsible for the maintenance of order while nurturing students to social and academic improvement. Given the importance of cooperation, the teaching role is highly dependent on teachers' authority. That authority in turn rests on complicated social dynamics. As Grant (1989) argued, the social basis of teachers' authority is the esteem accorded by the community to the role. Lortie's (1975) characterization of the status of teaching suggests the tentative nature of this social foundation of teachers' authority. Lortie found teaching to have a "special but shadowed" status, with individual teachers accorded less esteem

than the profession as a whole. (Lortie 1975). Grant later expressed concern that these sources of authority were eroding while external demands on teachers were increasing. (Grant 1989). This suggests that policies and public discourse that devalue teachers in the public mind run the risk of making it more difficult for teachers to do their jobs.

Motivation

Centralized accountability systems based on centrally chosen measurement instruments may impact motivation in three ways. The loss of discretion and autonomy that adds perceived value to teachers' work was introduced earlier when discussing resistance. Beyond that, legibility efforts based on accountability might also affect motivation by reducing their ability to extract value from non-monetary rewards and by devaluing goals that are not captured by the accountability system.

Teachers in general are more motivated by non-monetary rewards than other workers. (Ingersoll 2003). As with other SLBs (Lipsky 1980), teachers have a high public-service orientation and place a great deal of value on intrinsic rewards and the sense that they are making a difference. (Lortie 1969, Lortie 1975, Ingersoll 2003). At the same time, the teaching job does not make it easy for them to know whether they are doing a good job or not, as the results are uncertain and in many cases will never be known by the teacher (the benefit to students may not be apparent until they are adults, for example). (Ibid). They seek evidence of goal achievement in prideful occasions (e.g. spectacular cases, appreciative graduates) that are rare and generally unmeasured by the organization. This leads to a need for what Lortie (1975) calls "reassurance capital." Unfortunately, the sources of reassurance capital are limited for teachers. Validation through career advancement is limited by the flat nature of the profession. Moreover, the

structure of training and recruitment does not build self-esteem like it does in other professions. The process of socialization of teachers is weaker and less demanding than for other professions. The ease of entry deprives teachers of a sense of pride in having gotten in. There is therefore a greater emphasis on external validation from the public's valuation of teachers and teaching than in other professions.

The value they assign to goals distinct from those of the bureaucracy in which they work is related. For example, teachers attach significance to their work beyond the curriculum, setting as key goals instilling morals, inspiring a love of learning and including all students. (Lortie 1975). These goals are entirely excluded from the CCSS, evaluation systems and the instruments used to measure them (namely, PARCC and evaluation rubrics). To the extent that these goals are teachers' true motivation, reforms that minimize them risk reducing teachers' incentives.

Control and flexibility

Autonomy and discretion come into play beyond self-perception and motivation. They may be necessary elements of classroom technology. Teaching is highly variable, context dependent and judgment-intensive. It is "a complex and subtle craft" that does not lend itself to mechanization. (Sizer 1984 p.4). "That students differ may be inconvenient, but it is inescapable." (p.194). To be responsive to these differences, teachers may need to retain flexibility. Ingersoll (2003) suggests that poorly designed accountability systems may deny teachers the control and flexibility necessary to do the job effectively.

Collaboration

A final condition of success that simplifying accountability systems may threaten is collaboration amongst staff. As Perrow (1986) notes, the act of measuring workers can

lead to self-regarding behaviors and this can upset collaborative cultures. Such collaboration is important in education as it supports desirable behaviors such as sharing successful practices and materials.

2.2.4 Poor implementation

Implementation plays a role in the success of any policy. The extent to which a policy or program is implemented with fidelity is fundamental to program and policy evaluation. Successful implementation in turn depends highly on understanding local factors and context. Poor implementation is thus highly related to the conditions under which legibility efforts fail; it is difficult to successfully design and implement an intervention aimed at local actors while devaluing local knowledge and ignoring context. In the case from Scott (1998), for example, what looks like a failure to implement policing of the forests successfully can be traced to the failure to acknowledge the role forests played in the lives of local communities. Were authorities to take that into account, they may have designed the policy with at least some chance of being successfully implemented.

In education, there are very strong questions about whether schools as organizations and the administrators tasked with implementing accountability policies are situated to implement them effectively. Elmore (2000) argued that schools were not structured for standards based reform and principals, insofar as they are not trained as instructional leaders, were not equipped to implement the reforms: "Here, then, is the seeming conundrum: Schools are being asked by elected officials—policy leaders, if you will—to do things they are largely unequipped to do. School leaders are being asked to assume responsibilities they are largely unequipped to assume, and the risks and

consequences of failure are high for everyone, but especially high for children." (Elmore 2000 p.2). Chubb and Moe (1990) also raised questions about the structural capacity of education organizations to implement reforms, arguing that policies were destined to fail because they ignored the limits of the system charged with implementation. Ingersoll (2003) shared these concerns: "...many top-down school reforms betray a deep lack of understanding of teachers' work and the way schools actually operate." (p.235). They tend to "divert attention from organizational sources of school problems." (p.236). Thus, it is worth asking how well New Jersey's reforms were implemented in light of the high degree to which they devalue local knowledge and ignore context.

3 Research Questions and Methods

This dissertation relies on three case studies using both qualitative and quantitative methods. Qualitative methods, however, dominate, given the goal of uncovering hard-to-quantify underlying mechanisms at work in local contexts. The qualitative methods include in-depth interviews, participant observation and document review. Quantitatively, I relied primarily on descriptive statistics, though I include a logistic regression to evaluate turnover in one of the districts.

3.1 Research Questions

This dissertation poses the following research questions:

1. To what extent do recent education reforms meet the conditions that led to the failure of other legibility efforts?
    a. To what extent do the reforms involve the simplification of a complex social process?
    b. To what extent do the reforms shift expertise from local control and situated knowledge to centralized experts?
    c. To what extent do the reforms ignore context?
2. If the conditions are met, to what degree can they be linked to the failure of reforms?
3. What is the risk that reforms will lead to negative consequences beyond not improving student outcomes?

3.2 Case Selection

Evidence in this dissertation comes from three districts. I selected two purposefully for in-depth interviews of teachers and principals. I was an employee in the third, enabling participant observation and access to administrative data as part of my ordinary duties. The three districts effectively bookend the spectrum of NJ contexts, with one being wealthy and suburban and the other two lower-income and urban, allowing me to make reasonable inferences about generalizability. The argument for generalizability is twofold. First, if conditions exist at both extremes, they are likely to exist in between as well. Second, NJ is to an extent a state of extremes, with a large share of districts either relatively affluent in suburban settings or lower income in urban settings.

For the two districts in which I conducted in-depth interviews, I sought out one wealthy, racially/ethnically homogenous, suburban district (hereinafter district A) and a high poverty, racially and ethnically diverse, urban district (hereinafter district B).[8] For district A, I reached out to several such districts, relying largely on professors' introductions to pitch my research project. I selected the first district to approve the project. For district B, I again relied on professors' connections, and the first district I contacted approved this project. I did not select the third district for this project. Rather, I was awarded a fellowship and was accordingly placed in an urban New Jersey district that had requested a fellow (hereinafter district C). The district was in some ways a more extreme version of district B.

[8] For the sake of protecting participants, I mask the names of each of the three districts, and, to the extent possible, suppress details that might make identities obvious.

In the table below, I summarize the districts on several key characteristics. District Factor Groups (DFGs) offer a convenient, if highly oversimplified, shorthand for describing the relative socioeconomic status (SES) of districts. They were initially developed in 1975 to compare the performance of students in demographically similar districts on standardized tests. They have been updated several times since 1975 (with every national census) but retain the same basic structure. As with many summary measures, DFGs have been used for more than their original purpose, most notably to determine the initial group of districts that came to be know as Abbot Districts after the landmark Abbott v Burke case. The DFGs are based on six characteristics (high school graduation rates, post-secondary educational attainment rates, unemployment rates, occupational status, poverty rate and median family income). DFGs range from A to J, with A representing the lowest SES districts and J the highest. (New Jersey Department of Education n.d.).

While DFGs are a proxy for differences between districts, a more granular look at the data shows just how different the districts are in both conditions and performance. Regarding conditions, while district A is almost entirely funded by local property taxes, districts B and C are funded largely by the state. The state took over district C, selecting a superintendent and rendering the local school board advisory. In terms of size, none of the districts are particularly large, though district A is the smallest at less than half the number of students of district B. District C is the largest at around 18,000 students, but the districts large charter population means it directly served fewer than 10,000 students in 2017-18. Both districts B and C have higher student teacher ratios than district A and less experienced teachers. They also have dramatically larger shares of economically

disadvantaged students and English language learners (ELL students) than district A. (New Jersey School Performance Reports n.d.)[9]

Table 3.2 Case Selection

	District A	District B	District C
Relationship b/w District and State	Funded mostly by local property tax revenue	Highly dependent on the state for funding;	Highly dependent on the state for funding; state took over district
Enrollment	<4,000	>10,000	~18,000
District Factor Group	I	A	A
# of Schools	<10	>10	>20
# of "Focus" Schools (2017)	0	5	14
Student-teacher ratio	10:1	13:1	13:1
Average Teacher Experience (2017-18)	14.7 years	10.2 years	13.5
Economically Disadvantaged Students (2017-18)	3.6%	77%	56.3%
Students w/ Disabilities (2017-18)	17.9%	15.6%	15.3%
ELL students (2017-18)	0.7%	28.6%	10%
4-Year Grad Rate (2015,2018)	96.9%, 97.7%	68.5%, 73.4%	63.6%, 68.5%
AP/IB Participation % of 11th/12th (2017-18)	51.3%	26.6%	19.8%
College Enrollment (2017-18)	90.2%	60.1%	43.4%
Students Meeting or Exceeding Expectations ELA, Math (2017-18)	75.5%, 65.6%	31.1%, 23%	14.1%, <10%[10]
Chronic Absence Rate	11%	11.7%	30.2%

[9] To protect district identity, I cite to the general search page rather than the district's SPRs.

[10] To protect privacy, School Performance Reports do not show low percentages.

The performance differences are in some ways more dramatic. District A students are substantially more likely to meet or exceed expectations on state ELA and Math exams, take advanced courses, graduate high school and attend college. Regarding graduation, it is worth noting how stable the outcomes are. Districts B and C showed some improvement in on time graduation from the 2015 to 2018 cohorts, but not enough to fundamentally close the gap with district A. District B and C are not identical, however. While district B has a chronic absence rate in line with state averages, for example, district C's rates are roughly three times as high. District C students also attend college at substantially lower rates than district B. (New Jersey School Performance Reports n.d.)

While these districts differ significantly, they have at least one thing in common. They use the same evaluation measures. By virtue of being in New Jersey, all three districts use mSGP as the student growth measure and include SGOs in the overall summative rating. While they have a choice of observation instruments, all three districts use the same rubric as the California districts in Stecher et. al. (2018), Danielson's FFT, to evaluate teacher practice. FFT consists of 4 domains: Planning and Preparation, the Classroom Environment, Instruction and Professional Responsibilities. Each domain has five or six items and each of those has between two and five elements. In total there are over 60 distinct elements in the framework. When used as an instrument to evaluate teacher practice, administrators generally must find and note evidence for each of the 22 items, referencing the elements within them. (Danielson 2007).

3.3 Qualitative Methods

3.3.1 Qualitative Interviews

I conducted in-depth qualitative interviews of teachers and principals in grades 3-5 in districts A and B. I restricted the sample to grades 3-5 to ensure I captured the core "battleground" of reform - third grade is the first tested grade and all elementary teachers teach both math and ELA in these grades. This also simplifies the analysis, as it eliminates potential differences arising from different middle and high school experiences of reform. The tradeoff is that my findings may not be generalizable to higher grades. I accepted this tradeoff with the knowledge that further testing of the framework proposed here should include higher grades.

I selected participants primarily by snowball sampling, relying on my main point of contact in each district to recommend and connect me with possible interviewees. I reached out to each potential interviewee and schedule interviews with those who were willing to participate. In total, I conducted 19 interviews with 20 teachers – two teachers opted to interview together - and interviewed principals in 7 schools. The samples were relatively balanced between districts. In district A, I interviewed 9 teachers and 4 principals in 4 schools, along with the superintendent (on background only). In district B, I interviewed 11 teachers and 3 principals in 3 schools. In both districts, each participating school had at least a principal and one teacher volunteer.

Interviews took place across parts of the 2014-15 and 2015-16 school years in both districts. While I conducted many of the interviews in the fall semesters, some did take place in the spring. In analyzing the results I have made efforts to account for potential changes in perception and attitude at different times of year. Teachers

interviewed in March near PARCC for example, might be more stressed, with more acute negative perceptions, than teachers interviewed nearer the start of school. Despite my efforts, it is possible that interview timing may impact the results. Given that more of the interviews took place in the fall, however, and that my findings were generally consistent regardless of timing, the risk that the results would be meaningfully different had all interviews taken place in one season is low. I interviewed most teachers and principals only once and interviews generally lasted about an hour, with a few as short as 45 minutes. Some interviews, however, covered two or more hours across two interview dates.

The interview protocols for teachers and principals can be found in the Appendix. I designed the protocol as a broad guide to facilitate open-ended discussion of staff experiences of the key elements of reform – common core, standardized testing and performance evaluation. To minimize biasing staff responses and in the hope of drawing inferences about what was most salient in their experience from how they chose to respond, I framed the questions as general inquiries about each element (e.g. "Tell me about the teacher evaluation system" or "Tell me about the tests your students take") rather than directly asking my research questions ("Does this shift expertise"). While the protocol includes questions about how participants felt about the changes or asks them to characterize the changes, as a practical matter, I rarely had to ask this directly. Staff generally provided their opinions and characterizations in response to the broader initial questions.

Before diving into substantive questions about the elements of reform, I asked simple questions build rapport with the staff and give them an opportunity to talk about

themselves. This also allowed me to capture background information that informed the inferences I drew. For example, I asked teachers about how long they had taught and what subjects, easy questions for warm-up but also opportunities to test whether opinions were different based on subject and experience. Similarly, I asked all teachers what they thought their primary duties were as a teacher. This was designed as a rapport-builder but proved invaluable to my analysis of whether reforms accurately characterize the problem.

3.3.2 Participant Observation

As noted in section 3.2, my study of district C was opportunistic sampling that took place not by design but rather as a byproduct of a career opportunity. As a senior administrator in a high-poverty urban school district, I was able to observe firsthand many of the processes and interactions about which I was asking in districts A and B. In total, I spent 2.5 years in district C beginning in Fall 2015, with my main duties including financial planning. For 1.5 years of that, I also managed the central office staff responsible for implementing the teacher and principal evaluation systems.

Beyond the summary information provided in section 3.2, district C proved a rich case study for at least two reasons. First, as a state takeover district dependent on the state for the vast majority of its operating budget, the internal and external politics were always salient. All major decisions needed to account for a wide array of local and state stakeholders. Second, district C had a large, fast growing charter sector. This added complexity to decision-making and stress to interactions with many local stakeholders, including staff and community members. It also added another set of stakeholders to the mix, as external organizations with an interest in the fate of charter schools looked on. Both the state takeover and the charter growth are directly related to the subject of this

dissertation as both at least in part depend for their rational on district C's poor performance on standardized metrics.

Against this background, I experienced several events of note directly relating to summative evaluation. For example, I participated in the process by which principals selected performance measures for a portion of their own evaluations. The process included individual meetings with the principals and their central office supervisors, allowing me to witness directly how principals interacted with and responded to their central office supervisors. I likewise participated in several trainings of principals, including trainings on two tools my team created that are relevant here. One training was an all-school-leader training on a digital tool my team designed to help them keep track of the data on which their evaluations were based. For example, if a principal selected absenteeism as one of his or her metrics, the tool would include a routinely updated tally of their students' absenteeism progress. The other training was to help school leaders and other school staff utilize a simple Early Warning System (EWS) tool. This tool was formative; my team designed it at the request of central and school staff who wanted to make it easier for school staff to target interventions to students who were "off track." The final version allowed school staff to identify students meeting a set of user-defined characteristics, plan interventions and log progress. In both trainings, I or a member of my team, walked users through how to use the tools and answered questions. The fact that one tool was related to summative evaluation while the other was entirely formative allowed my to draw inferences from the ways staff responded.

My experience also included events that inform inferences I draw in this dissertation about how dependency on state funding impacts districts' experience of

reforms. Because my role included financial planning for the district as well as performance management, I was frequently involved in reporting district progress on both fronts to the state. As such, I witnessed directly the degree to which the state-district relationship impacted decision-making.

Despite this access, I detail only a small subset of my findings from district C in this dissertation. There are several reasons for this, but the primary reason is that it would be difficult to protect the identities of the district or staff If I were to provide this detail and I do not feel comfortable with that level of disclosure, having been an employee rather than a researcher. Implicit in this is my expectation that I would not have been granted nearly the level of access I was granted were I to have proposed it as part of a research project. As such, most of what I learned in district C serves as background, reinforcing or informing what I learned and describe expressly from district A and B interviews. There is, however, one significant exception. After I left the district, I requested and received de-identified teacher evaluation data through the standard research request protocol. This dataset forms the basis of the quantitative methods discussed in section 3.3.4.

3.3.3 Document Review

In addition to the interviews and participant observation, I reviewed documents related to New Jersey's teacher and principal evaluation systems. These include laws like TeachNJ, implementing regulations like AchieveNJ, and guidance issued by the New Jersey Department of Education (NJDOE) in support of these.

3.4 Quantitative Methods

As noted, quantitative methods comprise a small share of the analysis included here. Of the issues addressed in this dissertation, only teacher turnover lent itself to quantitative assessment given the nature of the issues and the data to which I had access. I relied on both descriptive statistics and a logistic regression model to assess the degree to which evaluation was associated with turnover and/or was changing the composition of the teacher workforce in district C.

4 Empirical Support: Reform Satisfies the Conditions of Failure

The framework proposed in this dissertation includes three conditions of failure: oversimplification of a complex natural and/or social process, a centralizing shift in expertise from those with situated knowledge to those with particular technical skills, and a failure to account for contextual factors that might impact the prospects of reform. This chapter details the high degree with which the core aspects of New Jersey's education reforms, particularly the accountability measures, meet all three conditions. For each condition, I address the extent to which reforms meet it based on (1) the legislation and regulations and (2) the practical experiences of the teachers and principals in districts A and B that are the primary subjects and objects of reform. While all three conditions are met in both districts, the degree to which they are met and the likely implications differ.

4.1 Education reform involves oversimplification of a complex process

The first condition of the failure of legibility efforts is the oversimplification of a complex process. That education is a complex social process is almost axiomatic. Stephens's (1967) description of education as a "complex and ancient process" better analogized to agriculture than to factory production or construction resonates. The question here is whether New Jersey's education reforms oversimplify education to such an extent that it can in part explain the disappointing results of similar reforms. More specifically, I ask to what degree New Jersey's reforms simplify the educational process and what technologies are used to effectuate that simplification. I start by reviewing legislation and regulations for evidence of abstraction, measurement and comparison.

Next I look for evidence of the same in the experiences of the teachers and principals that are the objects and subjects of the reforms. Through either lens, New Jersey's reforms involve exactly the type of oversimplification exemplified by the case studies in chapter 2.

4.1.1 On their face, New Jersey's education reforms simplify a complex process

On their face, New Jersey's education reforms oversimplify the complex process of educating children. The degree of simplification is dramatic, with a substantial gap between what is measured and what actually takes place. Likewise, the technologies by which this simplification is effectuated are precisely those that pose the greatest risk to complex processes: categorization of individuals for the purposes of sorting and making decisions about them.

TeachNJ, the 2012 legislation that required the creation of the current teacher evaluation system, not only mandates that teachers be categorized. It also specifies the four required categories -Highly Effective, Effective, Partially Effective and Ineffective - into which they should be placed.[11] It also prescribes the factors to be included in the summative measure that administrators must use to determine where to place them: a measure of teacher practice and multiple measures of student progress.[12] AchieveNJ, the

---

[11] "The State Board of Education shall promulgate regulations pursuant to the "Administrative Procedure Act," P.L.1968, c.410 (C.52:14B-1 et seq.), in accordance with an expeditious time frame, to set standards for the approval of evaluation rubrics for all teaching staff members, other than those included under the provisions of subsection b. of section 17 of P.L.2012, c.26 (C.18A:6-123). The standards at a minimum shall include: four defined annual rating categories: ineffective, partially effective, effective, and highly effective." (NJSA 18A:6-124.24),

[12] "Evaluation" means a process based on the individual's job description, professional standards and Statewide evaluation criteria that incorporates analysis of multiple measures of student progress and multiple data sources. Such evaluation shall include formal observations, as well as post conferences, conducted and prepared by an

implementing regulation, goes further, specifying the instruments, metrics, calculations and cut points that lead to placement in each of the four categories identified in TeachNJ. AchieveNJ established that student growth would be measured by Student Growth Percentiles ("SGPs") and Student Growth Objectives (SGOs"), and that instructional practice would be measured by three teacher observations according to a district-selected, NJDOE-approved rubric. (NJAC 6A:10-4). Principal practice would similarly be measured by observations and measures of student growth. (NJAC 6A:10-5).

The student growth measures are likewise highly simplifying. The simplification is to a degree what Rowan and Raudenbush (2016) referred to as distortion; the measures fail to capture the majority of the outcomes we want teachers to support. SGPs are a single number designed to characterize individual student's performance based exclusively on annual standardized tests in only two subjects. As mandated in AchieveNJ, SGP "means a specific metric for measuring individual student progress on Statewide assessments by tracking how much a student's test scores have changed relative to other students Statewide with similar scores in previous years." (NJAC 6A:10-1.2). The specific metric that NJ uses and how it is calculated is detailed in section 4.2 below.

Median SGP ("mSGP"), the metric that translates teachers' students' SGPs into a single score for each teacher is even more simplifying. For teachers that have at least 20 qualifying students within a school year or across two years, their evaluation includes a percentile score rescaled to a 4-point scale. The percentile score is simply the median of

---

individual employed in the district in a supervisory role and capacity and possessing a school administrator certificate, principal certificate, or supervisor certificate. (NJSA 18A:6-119.3).

all of their students' Math and/or ELA SGPs in that year (if they have 20 students with an SGP) and or that year and the preceding year (if they have fewer than 20 students within a given year). The score for any two teachers, therefore, can be the same even if all but 1 of their students have completely different scores. mSGP thus involves simplification across several parameters: it accounts only for performance on annual standardized tests covering at most two subjects, applies only to teachers with at least 20 students with SGPs across consecutive years – excluding, for example, all teachers in grades K-3 - and effectively ignores the performance of all but the middle student. It is worth noting here that principals and schools likewise receive mSGPs. In their case, it is the median score of all students with an SGP in the school and is therefore arguably even more simplifying than it is for an individual teacher.

SGOs are similarly simplifying. Chosen by the teacher, subject to principal approval, SGOs are a single metric based on the share of students meeting identified growth benchmarks on a single instrument. Teachers generally choose one or two SGOs, often in only one subject. Their rating is generally a conversion based on a highly simplified scale. For example, 75% of students may need to increase their score in a pre-post between the beginning and end of the year on a reading instrument such as DLM[13] for the teacher to receive a 3. A smaller percentage is a 2 and a higher percentage is a 4. SGOs thus generally exclude students outside the target percentage, all subjects other than the one chosen, and all aspects of that subject not captured by the chosen instrument. Principals have an analogue to the SGO but there is a broader range of options for what it includes.

[13] Dynamic Learning Maps, an alternative assessment for students with cognitive disabilities. See dynamiclearningmaps.org.

The instructional practice component of teacher and principal evaluation makes up the largest share of teacher evaluations and half of principal evaluations. They are also simplifying. As in the Improving Teaching Effectiveness districts, principals, assistant principals or subject supervisors conduct observations using a rubric, in this case FFT. Teachers' practice scores are aggregated from the results of three observations, at least two of which are scheduled in advance. These observations therefore represent a snapshot of what's happening in a classroom. The school year generally involves 180 school days. Even with the share of days devoted to testing and other non-instructional events, three observations capture only a very small share of what's happening in a classroom. Further, there is no requirement that any effort be made to ensure the three days are representative of daily practice and I found no evidence that any such efforts were made in any of the three districts studied here. As detailed in section 4.1.2, while staff in general found the observation rubric useful, their frustration with its use for summative purposes was based largely on the unrealistic simplification it entailed. As with SGO's, principal practice scores are based on rubrics that are far less prescriptive. Still, they receive a single score based on some distillation of their daily practice captured from a limited set of observations.

Each of the measures of educator performance and practice are therefore highly simplifying in their own right. However, there is one more simplifying step before ratings are final. Educators are ultimately sorted into one of the 4 categories by a single score that aggregates their scores on the individual metrics. While the state has changed the weighting periodically, most recently to reduce the weight of student growth measures, the aggregate score is always a weighted average of whichever scores each educator has.

For example, in 2016-17, the state increased the weight of the mSGP score. The score for a teacher who had a mSGP score would be weighted 30% mSGP, 20% SGO and 50% practice. For a teacher without a mSGP score, their score would be weighted 20% SGO and 80% practice. In 2018-19, the current administration reduced the weight of the mSGP score. The weights are now 5%, 25% and 70% for teachers with a mSGP score and 15% and 85% for teachers without a mSGP score.[14] Regardless of the ratios, the primary metric upon which educators are evaluated represents a simplified summary of three already highly simplified metrics.

The reforms anchored by TeachNJ and AchieveNJ therefore represent simplifications highly analogous to the case studies reviewed in chapter 2.1.1. Scott's scientific forestry case is the closest analogue. Whereas there, a highly complex ecosystem was reduced to a single number, in the case of NJ's reforms, the single number simplifies the highly complex social process of educating students.

4.1.2 The subjects and objects of reform experience the simplification

Teacher and principal experiences of the rating system reflect the oversimplification laid out above. They question the validity of their evaluations and express frustration with what does not count. Intuitively echoing the more formal discussion of distortion (Rowan, B. and Raudenbush, S. 2016), they argue either implicitly or explicitly that the metrics are too simplified to be valid measures of what is happening in the classroom. And they highlight in greater detail just how simplified the metrics are. This seemed to be consistent across districts, with staff in district's A and B reacting similarly to the exclusion of so much for the purpose of measuring their

[14] See state guidance here.

performance. It is worth noting, however, that principals in District A seemed more troubled by this and more protective of their teachers than District B principals, an issue I'll turn to in section 4.3.2.

As noted in the methodology section above and reflected in the protocol in the appendix, I did not explicitly ask any interviewees about simplification or any of the other themes in this dissertation. Rather, I asked them about their understanding of the evaluation system and invited them to reflect on it. As such, I am confident that the comments they made and from which I infer their experience of simplification are authentic.

District A Principals

All three principals in the subsurban district A made statements reflecting their experience of the simplification laid out in section 4.1.1. Their comments were wide-ranging but there were a few key ideas that emerged. First, they expressed frustration with what they perceived as invalid instruments to measure performance. From their elaboration, I infer that simplification is the root of their concern with validity: the instruments are too simple to capture what's really happening in the school. Second, they expressed broader frustration with the ratings and their calculation. From this I infer an intuitive discomfort with standardization and ranking.

Instrument validity was a prominent source of frustration for all principals. Regarding observations, one principal stated:

> Informal walkthroughs are way more *authentic*. In order to get a more accurate and fair assessment of what things look like in the classrooms and what teachers practices look like you gotta be experiencing it first hand. It's much less authentic than going into a classroom in a formal capacity. It's your informal walkthroughs and not with a tablet and taking notes for a Danielson model but simply because you wanna see what's happening in classrooms.

The use of the word authentic highlights the connection to simplification, as an overly simplified instrument cannot possibly capture the complexity of a classroom. The same principal also noted specifically how limiting it is not to be able to go "off script" in FFT when, for example, an item in the rubric does not apply to the type of lesson the teacher is doing. This also emerges repeatedly in teachers' reflection on observations. This principal used the word authentic again when describing an alternative instrument his staff uses to better evaluate students' progress, stating "we've developed benchmarks that are more authentic, a little more formative. It better accurately reflects what's happening in our classrooms." Another principle echoed the sentiment regarding observations' inability to authentically capture reality, reflecting, "I tell them that there's no possible way that in the 3 visits, one that's unannounced, two that are announced, I could possibly see 4's in every single aspect of Danielson."

This latter principal also expressed frustration with the general challenge of trying to accurately reflect the performance of an individual with a single number:

> It pissed me off a lot last year. There were certain teachers that I went in, and I know how they are because I'm in the classrooms all the time, and when I did their observation it did not come out to a 3.56. It came out to a 3.49. And other teachers, based on the lesson came out to a 3.58. And at the end of the year, when I looked at those scores, I mean I have two teachers right now I can think that were effective and not highly effective by .1 or .2 or something like that. And it made me mad. Because I believe they are highly effective. But because of the three lessons that I happened to see, or because of their scores, that they were honest on their SGO, scored this way or whatever, they didn't get the highly effective ranking that I believe they should have. And somebody else that I think is a good teacher should not have gotten as high a score as they did. It pissed me off so bad.

The third principal flagged a different concern with the overall limitations of simplifying for comparison, stating simply, "You can't do anything that's not standardized."

District A Teachers

District A teachers echoed many of the sentiments of their principals. Many challenged the validity of test-based evaluation, SGOs and observations and some implied a philosophical resistance to the idea of quantitative accountability. However, they expressed the latter slightly differently than their principals, perhaps reflecting their position as the primary target of TeachNJ.

Like principals, teachers' challenges to the validity of the instruments on which evaluation is based reflect an intuitive sense that the instruments are too simple to reflect reality. Said one teacher expressing a common theme: "One lesson, one test, cannot tell me what's good." Covering observations, the tests on which SGP is based, and SGOs, the teachers' contentions were familiar. Arguing that FFT is not valid as a summative instrument, one teacher asked rhetorically, "how do I show everything in 20 minutes?" She also noted that it is unreasonable for any observer to account for the many factors that impact a classroom on any given day, citing the example of a class in which one student's parent had passed away shortly before. Her colleague, a special education teacher, seemed to validate this limitation. She noted that one particular aspect of FFT requires students to lead or drive the instruction. She argued that students' disabilities rendered it functionally impossible for her to get a 4 on that without the observer "fudging" it. Similarly, a different teacher noted, like the principal, that while it's obvious that not all elements of FFT are relevant in all classes, the district did not allow NA's. Her example was the item requiring teachers to demonstrate proper correction of student behavior. If students were behaving so corrections weren't needed, she would receive a

low score, not an NA. This echoes Rowan and Raudenbush's (2016) flag about the nature of subject-agnostic rubrics like FFT.

Another teacher raised a slightly different issue with FFT that goes beyond observation. FFT's Domain 4, which addresses "Professional Responsibilities," requires teachers to provide artifacts as evidence. She had these artifacts but, as an untenured teacher whose submission was due early, didn't prioritize providing all the detail she could have. As a result she "got dinged" and ended up with a 3.49 (.01 below highly effective). She found it troubling that the difference between Effective and Highly Effective depended not on her practice, but on how much time she spent providing documentation of it. "If I had just added other stuff that I had I would have been highly effective."

Teachers similarly dismissed the possibility of any one test accurately reflecting their students' knowledge. Interestingly, this came from both ends of the student performance spectrum. A special education teacher noted that her students struggle with tests. Her non-special education colleague had the opposite sentiment: "Our students would do well whether we are here or not." Another teacher gave a concrete example of what the tests missed. "Two of my students finished PARCC in 5 minutes because they thought it was a videogame. One couldn't finish because he had meltdowns…" Other examples include the fact that the test is computer-based but doesn't account for different levels of keyboard skills. Echoing the principal, one teacher said she relied on other assessments, noting that PARCC "doesn't prove anything to a lot of teachers" demonstrating one of several coping mechanisms to which I will return in chapter 6.

Two broader challenges to the simplification of the test stood out. First, teachers noted that the administratively necessitated time-frame meant the test couldn't capture any learning that took place after the March administration of the test. Second, there was a sense that tests failed to capture too many things that they and their schools do. As one teacher detailed:

> I make the argument that we are stronger than what NJASK shows because of our sense of community, level of parental involvement, satisfaction, our ability to differentiate instruction. We really educate the whole child here. I've tried to incorporate elements of a private education in a public setting. [There is a h]uge commitment to service learning here. We do a lot to give back. It's a theme throughout the year. I worked really hard with parents and teachers to provide healthier options to kids in the cafeteria. We have a new food provider and an organic garden.

As shown below, this was a theme for District B as well, though perhaps with slightly different non-Math/ELA skills being taught.

SGOs might be the most limited measure, with teachers' comments suggesting they are simply a number for the sake of generating a number. As one teacher volunteered, "We chose SGOs we knew our students could perform well on" adding that even if they had not they could simply teach instrument-specific material the day before the students were tested. Another echoed this, saying the SGOs are purely administrative, forced and easily skewed, with no impact on instruction.

Like their principals, the teachers also raised philosophical arguments pointing to the simplifying nature of their evaluations. As one teacher put it, "it attaches a number to something that is bigger than a number." Another teacher seemed to channel Foucault, frustrated with "being labeled a number." Her colleague intuited the relationship between standardization and summative evaluation, noting that, because of time and capacity

constraints, summative evaluation cannot be done without standardization, and lamenting the corresponding impact on discretion (an issue to which I return in chapter 6).

While it is common to think that staff simply do not want to be evaluated, in this case, many of the teachers in both districts were sympathetic to or even welcomed being evaluated. That is why the evidence here is so compelling. Statement's such as "one lesson, one test, cannot tell me what's good," reflect not aversion to evaluation in general, but a desire to be evaluated fairly and accurately, and an intuitive understanding that no metric that summarizes the results of one test and/or a few lessons could be sufficient.

District B principals

Principals in the higher poverty, majority minority district B shared some of their suburban counterparts' implicit concern with simplification, but with less frustration. Generally, they were less critical of quantitative evaluation and less protective of their staff. The latter is something to which I turn in more detail in section 4.3.2. They were not without concerns, however. For example they acknowledged the validity issues raised by circumstances such as a student having a bad day on the day of the test: "If a kid on that day has a bad day, the one test will tell the wrong story." However, their response was more parsimonious. The principal who noted the issue when a "kid has a bad day" suggested only reducing the weight of the test. Likewise another principal noted that while sometimes the test is not a great measure of what the school is doing, "sometimes the test is a good measure of what we do." Their concerns were far stronger when it came to administrative burden than validity or philosophy.

District B Teachers

Unlike their principals, district B's teachers were at least as frustrated with simplification as their suburban counterparts. In fact, they were often more frustrated and were in some cases exasperated to the point of wanting to quit. Their responses and supporting explanations suggest the more intense responses may be a product of the very different context in which they work, including the different supports offered by their principals – something corroborated by their principals' less defensive posture. I will address this and other contextual differences in section 4.3.2. Here, I focus on those aspects of teachers' responses from which I inferred an experience of simplification. Like their district A counterparts, district B's teachers had validity-based issues spanning testing instruments, SGOs and the use of the observation rubric and philosophical issues with their evaluations.

District B teachers' validity concerns echoed those of district A, only more loudly. Like district A teachers, district B teachers were really concerned with what the observation rubric and tests did not capture. Regarding their observations, which were based on FFT, they flagged issues such as the failure to accommodate special education and the inability to use NAs for things that do not apply to the particular lesson being observed. Significantly, the word "authentic" came up here too, with a teacher arguing "observations are not really authentic, especially when they are announced… You can only show so much in 40 minutes." One particularly frustrated teacher called FFT "atrocious" and likened it to evaluating a chef based on one dish. "There's just so much. You're being judged after 2 or 3 visits on this big huge thing ... You can't have every food item in one dish." When probed, she acknowledged that FFT helps guide her

practice, suggesting her issue was not with the FFT model in general, but with its use as a summative instrument (a recurring theme throughout my interviews). District B teachers also added concerns about locus of control, noting that the teacher did not have agency over some of the things demanded by the rubric, though they did not specify which.

When it came to PARCC and its use in their evaluation, district B teachers' frustration was more pronounced than that of district A. They raised serious concerns about the validity of PARCC as an instrument to measure student learning and of mSGP as a measure of their own performance. One teacher used the word valid in nearly every sentence of our interview. She was not alone. Regarding PARCC as a testing instrument, staff challenged validity on several fronts. They argued that the test could not adjust for unique circumstances that impact students' performance on the day of the test. While the argument was similar to those made by district A teachers, the examples are telling of why the problem is likely more salient to district B teachers, reflecting the difference in context to which I will turn in section 4.3.2. As one teacher explained, "I have had a student that lived with grandma and mom came back in a week before the test and kicks grandma out with a restraining order. Do you think this student is going to pass that test?"

Teachers also argued that the test was capturing factors other than student's knowledge. There were several reasons for this, ranging from the mode of delivery, language barriers and cultural issues to the nature and complexity of the questions. Regarding the mode of delivery, teachers noted that computer based testing may capture more about students' comfort with the computer and digital tools than their content knowledge. As one teacher said, "kids have to learn how to use the test tools," giving examples of a digital protractor, log-in and equation editor. Another teacher presented

empirical evidence of the issue, noting, "I gave the practice test online and on paper. They did better on paper than online." This is more problematic in district B than A because "not all [students] have a computer at home." Likewise, language issues were more prevalent in district B, with a large ELL population, than in district A, where the vast majority of students were native English speakers. While there are some accommodations, one district B teacher put it succinctly: "More time doesn't help if [a student] can't read the question." The same could be said for cultural issues as some questions may presume knowledge that students in district B are less likely to have. One teacher showed me a question that required students to understand that being in a pool too long can lead to skin wrinkling, something that seems simple unless you do not have access to a swimming pool.

The nature and complexity of the questions drew the strongest reactions. One teacher called it "absurd" and "ridiculous" because the way the questions were asked was so challenging that students could not reliably display their knowledge. "You and I would have different answers yet the students are 10 years old. Whoever wrote questions is a whack-a-do." Her colleague provided more detail, arguing that PARCC is worse than NJASK at assessing students' knowledge of the standards because the questions were ambiguous or lacked necessary options. She pulled out a test and showed me an example of an ELA question in "A/B format."[15] Here the issue was that the question gave the students four choices "but those choices are not always really mutually exclusive. Or a clue that might help them isn't an option. So [students] have to choose one that isn't your

[15] In A/B format questions, students answer the question in Part A and in Part B explain the answer. In the example this teacher showed me, both answers were multiple choice.

option. This is not how to do this." Echoing the teacher that used the word validity in nearly every sentence, she went on:

> I'm already in the mindset that, regardless of how they do on the test, I'm not going to credit it with... I don't consider this test a valid assessment. No matter what the scores say, I don't have confidence in the validity of the test. This test has no validity. They have not established the integrity or validity of this test in any which way. So I don't feel it's going to be an accurate measure of what my kids know. I can trust that my kids have mastered this standard or are at least proficient in it. And this isn't the best tool to tell me that.

As the latter part of the quote highlights, teachers' issues with the manner and nature of testing reflect underlying issues with expertise. This is something to which I turn in the next section (4.2.1).

At least one teacher also flagged an issue with the inconsistency between PARCC, FFT and the curriculum.

> We are always differentiate, differentiate, differentiate, but all get same test. So much of the instruction we have to do is in groups. The curriculum is written and the way they want us to teach involves student collaboration. All day they work in groups then suddenly need to take a test by yourself and can't ask for help.

Finally, teachers argued that there was insufficient time to cover all material and that forced compromises in content. One teacher that was otherwise okay with the test said that it was not okay in March or May. Another teacher noted that they had had "tons of snow days" that year but "the test doesn't adjust for that either." Her colleague highlighted the implications. "We tried to throw as much as we could in there. But I feel like I kind of short changed my students because I was just going through things in like one or two days where in the past I've spent like 3 or 4 days on it." She was snapping at a high cadence while describing the rate at which she taught the material but sounded resigned and exasperated in describing her attitude toward the test. "We just gave it. Look. Try your best."

Regarding the use of PARCC in their evaluations (implicitly, regarding mSGP), they were highly skeptical that it accurately reflected their performance. Their concerns went beyond just basing their evaluations on a testing instrument they perceived as invalid. Some implicitly seemed to understand the limitations of mSGP as a summative measure of individual performance. "Historically my 5th grade students were on a 1st to 3rd grade level. Growth is tough to show in the way the state is expecting. Its sad." Another teacher echoed this, suggesting that even within her low performing district she had special challenges. "[T]hey put the lowest performing kids in my class. So I feel like I'm starting [at] a disadvantage, because [it's] much easier to bring up kids that are already reading on grade level." The language she used to describe how she felt - like "you're losing the battle" - was telling of the impact of evaluation on staff, something to which I turn in chapter 6. Here I note that these teachers' sense that mSGP was an invalid measure of their individual performance has support in the technical literature. (Betebenner, D. 2011).[16]

Like their district A counterparts, district B teachers were generally not against testing. Their issue was in the nature of the test and the manner in which it was being used. Comments like the following were common. "[I] don't have a problem with standardized tests, just it being the be all and end all," or "I kind of like the idea of the test. I like the accountability of needing to get to it. I just don't like how the test is used."

Like district A teachers, district B teachers were not nuanced in their take on SGOs. They flagged both the instruments on which they are based and the ease with

[16] As discussed further in section 5.4, mSGP is not valid for causal inferences and studies have found that it overestimates the contribution of teachers of high-performing students and understates that of teachers of low-performing students.

which they can be gamed as evidence that they do not capture anything meaningful. Like their district A counterparts, they ultimately regarded it as a wasteful administrative distraction. One teacher touched on several issues with it in one breath: "I'm gonna take a classroom of children that I have not evaluated for myself, then I will guess which children will make which progress, because [we] have to tier it in 3 levels. If [my] prediction [is] right, SGOs are okay. If not, SGOs [are] not okay. Somehow that's supposed to be indicative of what I taught." She also noted that with the instrument her school chose "it's different stuff at beginning and end of year. Why not use the same pre and post?" Another third grade teacher further detailed this validity issue:

> [We did] SGOs for the first time last year. [The] school came to conclusion we should use DRA. [17] Is it valid. No. Not a really good pre-post. All my kids need to read at 30 or above to be ready for 4th. But not going to set that as my goal for a kid that starts at a 6. That's kindergarten level. But DRA just made my SGO easier to achieve [than if it was my real goal of getting kids ready for 4th grade]. [We] wanted to use a pre-post. But the SGO they said to use … used second grade standards. That's stupid."

Other teachers directly addressed the issue of gaming SGOs by making them easy. "Problem I have is, what if someone has no integrity, they can put in any number they want." This is an issue she did not see as being resolved by comparison to other instruments, noting that you "can't compare SGO and PARCC. Apples and oranges." Echoing the "apples and oranges" analogy of her colleague, another teacher described SGOs in a way that highlights both that administrator approval is no guarantee of rigor and that many teachers understand the limits of non-standard evaluation methods: "We know its apples to oranges. [There is] nothing stopping me from choosing an 80 instead

---

[17] Developmental Reading Assessment, https://www.pearsonassessments.com/store/usassessments/en/Store/Professional-Assessments/Academic-Learning/Developmental-Reading-Assessment-%7C-Third-Edition/p/100001913.html

of a 90. We actually lowered it because the VP said [we] can drop it because 60 is all that is required. So [it's] not a reliable system yet. Definitely needs tweaks to make it so every NJ teacher is doing the same thing."

4.2 Education reform shifts expertise from local control and situated knowledge to centralized experts

The second condition of failed legibility efforts is a shift of expertise from locals with situated knowledge to centralized staff or those with specialized technical expertise. In the example from Bowker and Starr (1999), the government's efforts shifted expertise about individual's heritage from the individuals themselves to bureaucrats. In the example from Scott (1998), the government's efforts shifted expertise about the forest from trained forestry officials and those that occupied the forest to a central bureaucrat. This enabled a further shift that reduced the expertise necessary to operate the forest. Finally, in the example from Mitchell (2002), the government's effort to create a map of Egypt shifted expertise from locals with detailed knowledge of the historical, social, cultural and physical aspects of the areas in which they lived to an outside cartography expert and ultimately to the centralized bureaucrats who could rely on the resulting map. In this section, I review both the documentary evidence and interviews and find that New Jersey's education reforms shift expertise very much like the efforts in these three examples.

A qualifier may be helpful before detailing how NJ's education reforms shift expertise. The argument is not that shifting expertise is inherently bad. As with most services, pubic education necessarily involves a shift in expertise from, for example, parents to school boards, district staff and school staff. The question for this section is not

whether shifts occurred but the extent to which those shifts represent centralization of expertise and devaluation of situated knowledge.

Table 4.2: How did NJ reforms shift expertise?

From	To	Expertise About?	Summary
Teachers	Principals	Student learning	Principal has final say over the rigor of SGOs
	Private Company/ Psychometricians	Student learning	Student learning measured primarily by a test created by consortium run by Pearson
	State Bureaucrats/ Statisticians	Student Learning	Student growth required to be measured by SGP, which is calculated exclusively by DOE
	National Consortia	Standards; What students should learn	CCSS created by a national consortium led by the National Governor's Association Center for Best Practices
Principals	State Bureaucrats	Evaluation of teacher performance	DOE determines frequency of observations and which are announced and unannounced
District Administration	Private Company/ Psychometricians	Student learning	State-mandated test created by consortium run by Pearson
	State Legislature	Evaluation of teacher and principal performance;	Statute - defines elements of evaluations and scale on which educators will be measured; - specifies test-based growth measure
	State Legislature	Termination of school staff	Statute requires bringing tenure charges under certain conditions
	State Bureaucrats/ Statisticians	Evaluation of teacher performance	DOE: - determines share of evaluation score that is made up by each of the three elements; - calculates mSGP - for teachers w/ mSGP, calculates overall rating; - has oversight over choice of observation rubric
Boards of Education/ Parents/ Community	State Bureaucrats	Standards; What students should learn	Mandates use of CCSS

4.2.1 On their face, New Jersey's education reforms shift expertise from teachers, principals, districts and communities to legislators, bureaucrats, technical experts and private companies

As table 4.2 summarizes, New Jersey's education reform legislation and regulation explicitly shift expertise. The majority of these shifts involve devaluing local expertise in favor of the expertise of those further removed from what happens in the classroom. Moreover, the expertise is far from peripheral. Touching everything from what students should learn, whether they are learning and how to measure it to which teachers and principals are doing a good job and under what circumstances they should be terminated, New Jersey's reforms shift who gets to answer questions that are part of the very core of what is public education. On their face, then, New Jersey's education reforms meet the condition of shifting expertise away from those with situated knowledge towards those further removed.

From Teachers to State Bureaucrats and National Technical Experts

The most substantial shifts in expertise are from teachers to centralized bureaucrats and technical experts, including a private company. These shifts cover questions that hit at the core of the educator's role: What should students learn? Are they learning it? And how do we measure if they are learning it? Through this lens, New Jersey's education reform can be seen not just as an effort to hold teachers accountable. The accountability framework also fundamentally devalues the expertise of teachers and replaces it with that of different experts much farther from the classroom.

To National Experts: What Students Should be Learning

New Jersey's State Board of Education adopted the CCSSs in 2010. When they did so, they made the decision to outsource the decision about what students should learn

and by when to national experts. The CCSSs were developed by the National Governors Association and the Council of Chief State School Officers, who in turn relied on "educators, curriculum experts, school administrators and higher education faculty." [18] While teachers were incorporated into the design process for the CCSSs, they were drawn from a national sample and, within states, from a statewide sample.[19]

CCSSs identify "what students are expected to know and understand by the time they graduate from high school" and at the end of every grade level along the way. They are extremely detailed. For example, in third grade alone, the CCSSs cover 5 domains, each of which contains up to 9 standards.[20] While the CCSSs address only mathematics and ELA, as discussed in section 4.2.2, they impact other subjects; their breadth and depth combined with the timing of the testing of those standards sometimes limit teachers' choices about other subjects.

Notably, this was not an area that teachers were especially bothered by. As discussed further in section 4.2.2 and chapter 6, teachers were generally happy to have guidance about what they needed to cover and a structure that would help ensure their incoming students were consistently getting what they needed in prior grades. They were far more concerned about the devaluation of their expertise about whether their students were learning and how to measure it. This was reflected in their frustration with both SGP and the PARCC test on which it is based.

---

[18] https://www.state.nj.us/education/archive/sca/
[19] http://www.corestandards.org/about-the-standards/development-process/
[20] http://www.corestandards.org/about-the-standards/development-process/

To State Bureaucrats and Statisticians: Whether Students are Learning

The shift in expertise about whether students are learning from teachers to state bureaucrats and statisticians is especially analogous to the examples discussed in chapter 2. In particular, it strongly echoes the case of mapmaking in Egypt. Much like the Egyptian government enlisted outside experts using a highly technical procedure to build a "rigorous" map, New Jersey's government mandated a measure of student learning that required highly technical statistical expertise. Pursuant to TeachNJ, the primary measure of student learning is a student growth percentile. In general, student growth percentiles involve grouping students based on prior performance, then comparing students' current performance to their peers in their group. Comparing their performance does not necessarily require highly technical expertise. But grouping them does. (Betebenner 2011). The following is from a practitioner's guide to various growth measures and addresses the specific SGP model used in New Jersey:

> A strict implementation of this procedure would seem to involve the selection of "academic peers" that have identical previous scores. This is impractical and imprecise with large numbers of prior grade scores...The computation of SGPs involves a...statistical tool called quantile regression... Instead of fitting one line for the conditional average, the SGP model fits 99 lines, one for each conditional percentile, 1 through 99. As a point of reference, the 50th line is the line for the conditional median... This conditional median line represents the best guess about the median of an outcome given a predictor... (Castellano, K. and Ho, A. 2013).

Generally, teachers are not in a position to conduct this type of analysis. Here the developer of New Jersey's adopted growth model is Damien Betebenner, a two-Ph.D. academic who at the time of New Jersey's adoption of his model was at the National

Center for the Improvement of Educational Assessment (now the Center for Assessment). (Betebenner 2011).[21]

The shift of expertise away from teachers is not merely a product of the technical skills required to run the quantile regression. There are likely many statistics teachers in New Jersey who could do so. Teachers and individual school districts are also precluded from this analysis because it requires large datasets - large student numbers for comparison - that are generally available only at a statewide level. These datasets are the purview of state bureaucrats in DOE. (Castellano, K. and Ho, A. 2013).

There is an additional way in which the use of SGP as a measure of student learning shifts expertise. Like many technical policy solutions, the sophistication of the tool masks the normative decisions that are still required. Rather than eliminating judgment, these often involve shifts in what kind of judgments are needed and who gets to make them. For SGP, there are numerous technical judgment calls involved in the modeling. But a more fundamental judgment needs to be made to characterize the results of the model. To utilize SGP to make decisions requires making normative judgments about what constitutes an adequate SGP, which in turn requires normative judgments about a future goal and the time horizon to meet that goal. (Castellano, K. and Ho, A. 2013). By utilizing a complex technical measure, the participants in these judgement calls are limited. Only those that understand - or have the data to test - the implications of various different goals are in a position to make decisions about them.

There is a fair argument that, as a practical matter, this shift is of minimal consequence given how small a share of teachers' summative scores mSGP makes up

[21] https://www.nciea.org/about-us/team/consultant/damian-betebenner

(now at most 5% and for many 0%). However, as discussed below, teachers' experience of this shift was dramatic, suggesting that teachers, for various reasons, are more affected by the score than the numbers suggest they should be. That mSGP plays an outsized role in how educators talk about education reform can in part be explained by just how much educators value being treated as experts. mSGP based on PARCC is in some way the most direct threat to that. In light of that, the difference between perception and scoring is not as surprising.

To a Private Company and Psychometricians: How to Measure Whether Students are Learning

The above understates the shift in expertise implied by reliance on a measure like SGP. SGP requires not only technical expertise and statewide data, it also requires a "set of psychometrically sound tests over two or more grade levels in a single domain..." (Castellano, K. and Ho, A. 2013). To meet this requirement while aligning with the CCSSs, New Jersey adopted PARCC[22] tests for ELA and Math. This test involves two major shifts in expertise. First, it is created and run by a consortium managed by Pearson, a multinational headquartered in London that is the largest education company in the world. Second, it has to be scored centrally. School staff are not in any way involved in the scoring or review of the assessment itself. As I detail in section 4.2.2, this lens provides a different way to interpret teachers' near universal hatred of PARCC.

From Principals to State Bureaucrats

AchieveNJ also shifts expertise from principles to state bureaucrats by prescribing the frequency and nature of teacher observations. For example, principals are required to

[22] In 2019, Governor Murphy renamed PARCC as New Jersey Student Learning Assessment (NJSLA) and revised the test to have fewer questions and take less time.

observe tenured teachers three times, with one of those three being unannounced. This may seem a fairly simple mandate, but as described in sections 4.1.2 and 4.2.2, this is an area about which principals spoke a lot and from which we can infer a sense that this is an important part of their discretion.

From District Administrators to Legislators, State Bureaucrats, Technical Experts and a Private Company

Along with teachers, district administrators may have lost the most discretion under reform. The law and implementing regulations touch on major decisions previously within the discretion of district leadership. These include how to measure student learning, how to evaluate teacher and principal performance and, under certain circumstances, who they must terminate.

The adoption of the CCSSs and the mandate that student learning be measured using PARCC as the instrument and SGP as the metric was addressed in the section on teacher expertise. Here, I note simply that this was an expertise previously shared between district leadership and teachers and therefore represents a loss of discretion and a shift in expertise from both. While district leaders retain the authority to utilize other testing instruments and act on data other than SGP, as will be discussed in chapter 6, this is de facto limited by the practical considerations of implementing a test like PARCC along with the leverage mandatory standardized testing gives external stakeholders.

The evaluation of staff involves a more explicit shift in expertise. As noted in section 4.1.1, the tenure reform act specifies the elements of teacher and principal evaluations. All teachers and principals must receive a practice score. For teachers this must be based on observations using a rubric and, as noted above, these observations

must take place a prescribed number of times. While districts retain the freedom to choose their own observation tool, their choice is subject to DOE approval. In practice, most districts, including all three subjects of this dissertation, choose FFT. Both must also have a student growth outcome component. Teachers in grades four[23] and up with at least 20 students taking state standardized mathematics and ELA tests must have both SGO scores and an mSGP score. Those without 20 student test scores have only an SGO score. Districts and principals retain a fair amount of discretion over SGOs but not over mSGP. mSGP is calculated by DOE. (NJAC 6A:10-4.2(d)) Where teachers have a mSGP score, DOE also calculates their full summative rating - the rating aggregating all three scores .[24] (NJAC 6A:10-4.2(d)2)).

If the purpose of the final evaluation score were to be merely advisory or to be used as a formative data point, the shift in expertise about instructional staff evaluation might be less problematic. However, the legislature also mandated that the resulting score be used to make retention decisions under certain conditions. Pursuant to C.18A:6-17.3, districts must file a tenure charge - an action to terminate a tenured teacher - for all teachers that receive an ineffective rating the year after receiving a rating of either ineffective or partially effective. While there are no mandates for teachers receiving a partially effective rating, the district must provide written evidence of exceptional circumstances if they choose not to file a tenure charge for teachers receiving this rating the year after receiving any rating below effective.[25] In other words, the state has not only

[23] Because third grade students have no prior test scores, the state does not generate an SGP for them or an mSGP for their teachers.

[24] The time it takes DOE to produce the final summative ratings for these teachers has the strange consequence that decisions about staff often need to be made before the final summative score is known.

[25] https://www.state.nj.us/education/AchieveNJ/implementation/legalrequirements.pdf

taken the decision about which teachers are not fit to continue teaching out of the hands of the district, they have also decided which teachers are so poor that only exceptional circumstances can keep them in the classroom.

As a practical matter, districts can and do work to preserve some of their own discretion in the face of these mandates. However, as discussed in chapter 6, they do so as de facto acts of resistance. The effectiveness of their resistance depends a lot on context - see section 4.3 - with district A functionally retaining more discretion than district B.

From School Boards, Parents and Community Members to State Bureaucrats

I have so far considered the impact of statewide adoption of the CCSSs on educators and districts. However, there is also a shift away from school boards and, by extension, parents and community members. However indirectly, these are the stakeholders with the most interest in what students should learn and by when. The state board of education has effectively determined that communities cannot choose for themselves a purpose of their public educational institutions that does not include the purposes for which the CCSSs were designed. Whatever the merits of the CCSSs or the conflicting decisions local communities might have made in its absence, the loss of this choice can be a substantial one. The degree to which this represents a substantial loss of discretion depends on context. District A functionally retains more authority to prioritize as the community sees fit than does district B because it is less dependent on the state for funding and it's students relatively high performance ensures it is subject to less state scrutiny.

4.2.2 Staff experience the devaluation of their expertise

Section 4.1.2 highlighted how looking at the problems educators have with components of NJ's education reforms through the lens of oversimplification shows them to be more than self-serving complaints. This section highlights a similar point looking at their response to reform through the lens of the shift in expertise from them and their districts to the state and others further removed from the students. In both districts A and B, many of the teachers' issues with PARCC and their evaluations come down to their sense that their expertise is being devalued in favor of those of whose expertise they are highly skeptical. Principals, for the most part, did not raise many expertise-related issues. Table 4.2 suggests a potential reason for this: the number and importance of questions on which expertise is being shifted from teachers is significantly greater than that for principals.

As with section 4.1.2, I organize staff responses by district to identify potential differences in context that I will address in section 4.3. I retain separate sections for principals and teachers, but only to highlight the relative lack of expertise-related issues raised by principals.

District A Principals

Neither district A nor district B principals' responses suggested they feel particularly threatened in their core expertise. Table 4.2 shows that the primary area of discretion lost by principals surrounds how they observe and evaluate teachers. This is reflected in their comments discussed in section 4.1.2, which can be seen as problems with both how prescribed observations oversimplify the evaluation of teachers and how they challenge principals' expertise. One principal's comment that "to get a more accurate and fair assessment of what things look like in the classrooms and what

teacher's practices look like you gotta be experiencing it first hand" and that you "sometimes need to go off script" with FFT is a key example of this. The same is true for the principal that was "pissed... off a lot" because teachers "I believe...are highly effective...because I am in the classrooms" received final scores of effective. Both statements reflect the sense that they are expert in evaluating their staff and the system created by other experts may not reflect that expertise. Beyond this, there was a passing reference by one principal to the fact that expertise in assessing student learning should remain with the teacher: "The teacher can determine mastery in a way a test cannot." Overall, expertise shifts were a small part of principals' commentary on education reform. It was one of the dominant themes for teachers, however.

District A Teachers

Unlike their principals, district A teachers were aligned with their district B counterparts in their perception that they were better situated to evaluate their students' knowledge than remote test-writers and statisticians. Their responses ranged from explicitly addressing expertise to statements from which I inferred issues with shifting expertise. Regarding all statements addressing expertise, I note that my interview protocol did not include any direct prompts about expertise. I asked questions about expertise only as a probe or follow-up when teachers raised the issue.

Direct lamentations about the devaluation of their expertise were nevertheless common. One teacher hit it head on: "I'm not really treated like an expert. No one is saying we're the experts, we're the professionals," adding that as a result she felt "disrespected." She offered an analogy to drive home her frustration: "I fly a lot but I'm not gonna be a flight attendant." Another was equally direct, reflecting on how "expertise

is being centralized." In similar vein, her colleague offered that every teacher feels the devaluation of their expertise, especially when "the state is coming down with things." Only slightly less direct were complaints about their lack of involvement in the process of making the decisions that went into reform. As one teacher put it, "[p]eople don't feel like they have a say. People don't feel like their voices are heard." Another echoed this, stating, "If teachers participated in the creation of the rubrics we'd be more okay with it."

In addition to direct discussion of expertise, I inferred that teachers negatively experienced their loss of expertise from their explanations of why they disliked testing and their evaluations. Their comments suggest they have greater issue when it comes to evaluating performance than the adoption of CCSSs.[26] Teachers issues with PARCC and expertise were tied to their issues with its validity. They argue that the test cannot accurately measure student learning because it is a flawed instrument delivered in a particular mode on a particular day. Teachers, the argument goes, can better measure student learning because they are with the students every day and are trained to assess their knowledge. One teacher captured her colleagues' sentiments well when she said:

> How do you measure [unique personal experiences]? I could tell you any kid any day and tell you exactly how every kid is doing. But someone who doesn't know teaching but is looking at it from a far, will question my judgment... [The test result] doesn't convey that child. Children are not numbers. I could show them really well if I was designing it."

A special education teacher echoed this sentiment almost exactly:

> ...I have an advantage. I've had several of these kids for years. I see the growth from 2 years ago to now and its huge. One lesson, one test, cannot tell me it's not good. I'm not one of the teachers that asks admins what it takes to be a 4. Feel like I've been doing this for long enough that if I'm not there now I won't be. I'm putting everything into this while I'm here and home. One lesson, certain pieces, no background on kids...I know what I need to do.

[26] This is likely because they see formative value in the CCSSs that they do not see in the summative assessments. This is discussed further in chapter 5.

Other teachers noted how they work to insert their own expertise. One indicated how she downplays PARCC, stating "I use internal evaluations. The test doesn't prove anything for a lot of teachers." Another said, "[r]egardless of numbers I'm going to do what I think is best. I know I'm doing well from parent feedback and how students perform every day."

Teachers' issues with test-based evaluation are connected not just to their perceptions of outside experts' content knowledge but also and perhaps primarily to the delocalization of expertise. This is reflected in the geographic and/or personal language they use. The first quote above contains a clear example: "Someone who doesn't know teaching and is looking at it from afar will question my judgment." Other teachers lamented "[s]omone from *the outside looking in*" and that "someone that I don't know that doesn't know me or my kids is scoring me." Teachers' intimate experiences of students' performance based on daily interactions is in many ways the epitome of situated knowledge. Teachers' issues with PARCC and SGP reflect their experience of the degree to which that situated knowledge has been replaced with removed expertise.

District B Principals

As with their response to simplification, district B principals had less of an issue with any loss of expertise than their district A counterparts. I have several hypotheses for this difference. One hypothesis is that this is a product of a different relationship between teachers and administrators in district B than in district A. It might also reflect the possibility that the choices made by the legislature are more aligned to how district B principals would exercise their expertise anyway. Or it might be that district B principals, selected by administrators subject to the pressures of poor perceived performance and

financial dependency on the state, are more comfortable with the system into which they partially self-selected. These hypotheses are all related and are difficult to evaluate directly. However, in section 4.3.2 I address the extent to which evidence about differences in context is consistent with these ideas.

District B Teachers

As noted, district A and B teachers are aligned in their issues with how expertise is shifting. They address expertise explicitly and implicitly and challenge state and technical experts' content knowledge and where they are situated. However, there are a few differences. First, district B teachers were more concerned with the content knowledge of outsiders than district A teachers. Second, they were especially concerned with how appropriate the test was for their students' abilities or starting points, something that was not high on the radar of district A teachers. District B teachers found evidence of experts' inadequate content knowledge in what they perceived to be unreasonable expectations of their students. Third, district B teachers were more likely to include district administration and their own principals in the group of those whose expertise they challenged. Finally, as with their issues with oversimplification, their degree of frustration was greater; they seem to experience their loss of expertise more viscerally than their district A counterparts. Again, this may reflect the differences in discussed further in section 4.3.

Like teachers in district A, several teachers in district B offered direct and indirect assessments of expertise and challenged their exclusion from the process. One particularly frustrated teacher lamented that there are so many more tests now because "someone who has never been in classroom decided it would be good. People asking us

have no idea what they are asking us to do. What is this nonsense? Our expertise is not trusted. [Our] opinion [is] not valued in any shape or form." Another put it similarly, saying, "they don't allow the quote unquote expert - who would be me - to sit with you and help design it... We're professionals, we're colleagues. But you don't know what really goes on in that classroom...[L]et me decide how to allocate the time." Both referred to state and district officials in their statements. Their issue with their lack of voice was a common one, with other statements like "I have a voice that needs to be heard" arising repeatedly. Other teachers highlighted the advantage of their situated knowledge: "I know what my children will understand." One spoke for her colleagues generally, stating "I feel that - it's not just me, we talk about this all the time - we feel that the teachers in the classroom, we know what the students are capable of." Perhaps the most pointed assessment came courtesy of an analogy:

> But they don't ask us. We are the soldiers in the trenches. You have the generals up in that building who have no idea what combat is like but they are the ones making all the decisions for you and nobody listens and says, wait a minute, they're not ready for this. Or this is going too fast.

Another teacher referred to state officials and technical experts while noting that PARCC validity for her is an issue of expertise and local knowledge. She thinks the people who decided the test was appropriate don't know ten-year-olds as well as they should, adding that the "best person to know how good my kids are is me." She took a direct swipe at the academics she perceived to be responsible for writing the PARCC assessment: "I think that we are so busy creating assessments in academia without talking to kids and teachers that we are creating assessments that sound wonderful on paper" but bear no connection to ten-year-olds' reality. Her colleague joined her in the interview and laughed exasperatedly while she talked. The colleague agreed that test-writers lack

sufficient expertise. In her opinion it was because they "have not really worked with children" and therefore have no idea what a ten-year-old should be able to do.

Other teachers echoed their district A counterparts in showing how they work to preserve their own expertise in the face of pressure to abandon it: "I could care less [about my score]. I know what's going on in here...And that's why I shut my door. The state can say what they want but I know what they need. I see what they need."

4.3 New Jersey's education reforms ignore context

The final condition of legibility failure is ignorance of context. The examples in chapter 2 suggest that it is nearly impossible to simplify a complex process without ignoring context. Scientific forestry could not have been made so simple were it to try to account for the local residents' use of the forest. Aid agencies would have struggled more to identify simple solutions to the food shortage had they acknowledged the social context driving lower-income residents to have more children while wealthier residents placed disproportionate demands on production. And of course the accommodation of context in a system designed to categorize all people into a small number of racial groups would have made such a system far more difficult to administer. It is almost axiomatic that simplification and context are inconsistent with each other.

Here, I lay the foundation for further discussion of the implications of ignoring context in chapter 6 by highlighting some of the key differences in context that are not addressed by New Jersey education reforms and suggesting how those differences might explain some of the different reactions to reform seen between educators in districts A and B. While I focus on differences between districts A and B, differences can persist *within* districts as well. District B, for example, has one school situated in a suburb-like

setting despite being in an otherwise entirely urban district. If context matters, this suburban-esque school should experience reform almost like a hybrid of districts A and B. That appears to be exactly the case.

4.3.1 On their face, New Jersey's education reforms ignore context

In general, the evidence that New Jersey's education reforms ignore context is the absence of any language in the laws or regulations explicitly differentiating application to account for local differences. Without these exceptions, it is enough to note that relevant local differences exist that might interact with reform efforts. Section 4.3 highlights some substantial differences between the settings of districts A and B and the populations they serve. Section 4.3.2 addresses additional differences and how key differences affect the way teachers and principals experience the state mandates that make up New Jersey's education reforms.

Some argue that SGP and mSGP implicitly account for at least one key component of context by adjusting expectations based on each student's starting point. While this is technically true, there is growing evidence to suggest that even SGP and mSGP are sensitive to differences in context. In fact, they may be sensitive to exactly the context - different starting points of different students - for which they are designed to account. For example, Castellano and McCaffrey (2017) found that errors in commonly used aggregate growth measures, including the mSGP model New Jersey uses, are correlated with students' prior achievement. This leads to mSGP systematically underestimating the performance of teachers with students with low prior performance and overestimating that of teachers with students with high prior performance. (Castellano, K. & McCaffrey, D. 2017). The issue may reflect the inability of mSGP to

capture compositional effects such as that from the increased difficulty of educating 25 students who are all behind to differing degrees versus one student who is behind in a class where the other 24 students are on grade level.[27] As discussed further below, this is precisely the context with which teachers in district B and district A, respectively, are faced.

4.3.2 Staff experience the ignorance of context

Both district A and B teachers seem to be aware of the role context plays in their work and in the differences in context between their respective districts. Several staff worked in districts of both types - including one teacher who had taught in both district A and B. But even those who had not necessarily had experience in the other type of district acknowledged how substantial the differences are. While not all differences contributed to a more favorable experience of reform for district A staff - they experience far more parental pressure for example - on balance district A staff were far less likely to experience meaningful pressure to change their behavior as a result of reform. They were more insulated than were their district B counterparts. For district B staff, reforms were salient in their daily experience.

[27] It is also possible that SGP fails to account for summer learning changes. Summer learning loss is a well-established phenomenon that is far more common and dramatic for students in low-performing districts. If two students have similar prior scores but one makes gains over the summer while the other retreats, the SGP and potential mSGP may reflect that difference even though the teacher played no role in the difference.

Table 4.3a Summary of context differences between districts A and B

Area	District A Context	District B Context
Setting	Schools are situated in safe, residential, wooded suburban neighborhoods; there are no security guards	With one exception, schools are situated in low-income urban neighborhoods; visitors are greeted by security staff
Student Needs	Students generally have limited extra needs, with high prior performance and only rare issues outside of school	The majority of students have a lot of extra needs, with low prior performance and myriad issues outside the classroom
Parent Involvement	Parents are highly involved with high expectations, placing a lot of pressure of teachers and administrators; most are able to support students academically at home.	Parents are less involved with limited expectations of school staff; many are ESL immigrants unable to support students academically at home
Relationships b/w Teachers and Administrators	Generally positive; administrators are protective of teachers; teachers trust and respect administration	Far less positive; administrators are not generally protective of teachers; teachers are distrustful of central administration in particular
Resources	Highly sought after district allows them to be more selective about staff; very few first year teachers; nearly all classes have fewer than 20 students; PTO fundraises; able to offer extra-curriculars like orchestra; families generally provide learning opportunities over the summer	Staff are hired late, limiting the ability to be selective; new teachers are common; every teacher interviewed had more than 20 students; learning opportunities outside school and over the summer are limited
Relationship b/w District and State	Funded mostly by property tax revenue with limited dependency on the state; virtually no risk of aggressive state intervention	Highly dependent on the state for funding; seemingly perpetual risk of aggressive state involvement

Table 4.3a summarizes the key differences in context between the two districts. The differences in context span student needs, parent involvement, resources, the relationship between teachers and administrators, and the relationship between the district and the state. These differences provide reasonable explanations for the differences in

how district B staff responded during interviews. They can explain both the higher levels of frustration and the different focus of their ire. For example, district B teachers' students are much lower performing than district A students. Combining that with their sense that PARCC, SGP and mSGP don't account for this, a sense that is supported by research as noted above, their greater frustration with testing and their focus on reasonable expectations of students is not surprising. While all differences play a role in the experience of reform, student needs, the relationship between teachers and administrators and the relationship between district and state seem to dominate. I discuss these at greater length in this section.

Setting

I visited most of the elementary schools in both districts to conduct on-site interviews. I therefore had the opportunity to observe the settings in which these schools were situated. There are clear differences. District A's schools are situated in residential, wooded suburbs reminiscent of photos associated with the American dream. While I did have to go to the main office to sign in to all buildings, I did not encounter any security staff. In contrast, district B's schools are, for the most part, situated in traditional urban settings, dominated by the built environment with a mix of residential and commercial structures and limited green spaces. Several of the buildings were a sidewalk removed from major thoroughfares. And I encountered security staff at all buildings. One such encounter serves to drive home the differences between the two districts' settings. When an interview in district B ended after dark, I was set to walk to my car two blocks from the school. The security guard insisted on accompanying me. For reference, I am a six-

foot-two, 200-plus pound former college athlete. He walked me the two blocks to my car, parked in a neighborhood in which many of the schools' students lived.

One school in district B illustrates the impact of setting independent of other factors impacting districts' experiences of reform. Creating something of a natural experiment, one school in the otherwise all urban district B is situated in a classically suburban setting. It's surrounded by ample green space in an exclusively residential neighborhood. And as a teacher relayed to me, while it used to serve a less-diverse, higher-income population, that has changed dramatically in recent years. Nevertheless, expectations for this school are significantly higher than for the other schools in the district. As the teacher noted, "because you still have the culture of what its been and how it is [its suburban appearance], holds it above the rest." The setting thus appears to impact experiences independent of the other factors with which it is highly correlated.

Student Needs

The needs of students in district A are substantially different than those in district B. Unsurprisingly, given reform's focus on holding teachers' accountable for student outcomes, this may be the difference that most impacts the staffs' experience of reforms. While their higher performing students don't necessarily free them of all negative interaction with reform - see Parent Involvement below - even district A staff acknowledge how different their role is. Staff in both districts seems to agree that the role of teachers in districts like district B is far more complex owing largely to the differences in student needs. Again, my interview protocol did not include any questions about student needs. Interviewees volunteered all comments in response to questions about reforms.

District A staff had little to say about their students other than to comment on how good they were. As one principal said, “we have good kids here, they do what they are supposed to.” One teacher seemed to capture the sentiment best: "Our students would do well whether we are here or not."

District B did not echo that sentiment for their own students, though they seemed to recognize that places like district A do not have the same challenges. Instead, they discussed, sometimes at length, the academic starting points and out-of-school experiences of their students and their perception that reforms drastically underestimate these factors. Said one principal, “[you] can’t expect kids who babysit when their parents have two jobs or don’t speak English [to not be impacted]. We only have 6 hours with the kids.” Teachers agreed, with nearly all identifying both that their students come in behind and they face serious out-of-school challenges that cannot be ignored when evaluating them. Table 4.3b lists a sample of statements to this effect.

Table 4.3b: District B teachers’ statements about their students’ needs

“Only 3 of my students are on grade level. 100% are Hispanic. I'm an educator but also a caretaker. Basic needs need to be met first.”
“[We] spend a lot of time putting fires out. If I can generalize, in a district like this, you always have a lot of balls in the air. DYFS issues continue. Everything else. Teacher coverage etc. the stuff that eats your day up. It's a big factor.”
“Not all kids have a computer at home. I have had a student that lived with grandma and mom came back in a week before the test and kicks grandma out with a restraining order. Do you think this student is going to pass that test?”
“The children in this district don’t come with the basics. They’re at a disadvantage.”
“[We’re] constantly playing catch up. [We] spend [the] first marking period getting caught up. [The job is] more enriching but harder. Academic is the primary [goal] but often don’t get to that. Sometimes [students] haven’t eaten breakfast.”
“I have 23 students who probably would be ESL if eligible. Also parents don’t read to kids, don’t talk to kids, not really around, threaten [them] (not physically necessarily). Our kids parents don't even read or write. Our kids are learning letters whereas others can already read. Gets worse over time. Can imagine how overwhelming [that is] for a young child.”

Two teachers explicitly addressed how all these student differences relate to their roles and accountability. The first formerly taught in district A and echoed her district A counterpart's sentiment that the students need little extra support: "I used to joke when I was in [district A] that the kids would be fine if I played soft music. I never felt that I could take credit. Here the teachers are killing themselves. It's soul crushing." The other went after the state's failure to acknowledge the role of this: "[The] state needs to see it's not for lack of teaching here. We have to differentiate more. They don't differentiate in [a nearby wealthy, suburban district]. We have to show that we differentiate. [Differentiation is] not an option here." In short, teachers in district B experience the reforms differently and as more unfair than their counterparts in district A because they see student needs as key to understanding their role, and reform ignores that.

Parent Involvement

Parent involvement is a bit more nuanced than setting and student needs. As implied by some of the statements above, parents in district A are more able to support their students academically than those in district B and implicitly have a role in the fact that their students do not enter school with the same out-of-school challenges. They are also highly involved in supporting the schools directly, by, for example, fundraising through the PTO. On the other hand, they are far more demanding of schools and teachers and have used the data that is available as a result of New Jersey's reforms to amplify the pressure they place on their schools. As teachers and principals in district B tell it, their students' parents are less involved, less able to support their students academically and less supportive of the schools. However, they are also more deferential to teachers and do not add additional pressure. Nevertheless, my impression from the

sentiments of both district A and B teachers and principals is that parental involvement is a net benefit, even when they exert pressure.

In discussing the pressure they get from parents, one district A teacher noted "parent influence is so strong that [the] principal sometimes uses the fact that he is a parent to exert more pressure than [he can] as a principal." Others noted that parent pressure is felt primarily by teachers. In response to a question about what she sees as her primary duty, one teacher said "managing expectations for parents. For HR, [the] teacher dealing with parents is primary." Said another teacher: "In our district, parents pressure drives admin decisions. Admins struggle to reason with parents. Teachers are left to explain things." Seemingly adding a data point to the argument about administrative pressure, a special education teacher said, "School performance is published. [A] whole big reform came because SPED students did badly on NJASK." Still another talked about the nearly day-to-day involvement of parents in her class. "I just need to do some tests that we use as evidence for parents when they want to know why their student got a [particular grade on their report card]."

There is one story that I heard from several teachers and several principals as well as the superintendent that drives home how the data that reforms have made available to parents has increased the pressure they put on their students' schools. The story is of a concerned citizen analyzing state test data and identifying a pattern whereby one of the district's elementary schools underperforms the others. All interviewees were quick to point out that even the underperforming school is a high performer by state standards. Nevertheless, this concerned citizen's analysis caught the attention of parents more generally and they demanded action. At this point the telling of the story diverges, but

some contended that parents' demands were met with a push for more alignment across schools. Regardless of whether that action was taken as a result of parental pressure, it is clear from the number of staff aware of this that the pressure was felt widely. It is also clear that while this desire to compare schools existed before the reforms, reforms provided the standardized data that enabled more intense arguments from parents.

Another area in which parental involvement plays a role in district A and which bears directly on reforms relates to PARCC is opt-outs. A nationwide issue, opt-outs involve parents refusing to let their students take PARCC. Because the validity of the evaluation system depends on large numbers of representative students taking the test, opt-outs, especially from distinct subsets of the overall student population, pose a threat to the reform infrastructure. District A seemed to have a pretty substantial parent-driven opt-out movement, with one teacher calling PARCC a "mess" because of "tons" of opt outs. Another teacher said that many parents were refusing to let their students take PARCC, especially in high school. One spoke from experience, noting that she had three or four of her seventeen students opt out.

Teachers and principals also acknowledged the benefits of parent involvement in district A, with one principal noting "we have great parent support and involvement." A teacher was more specific, reflecting positively on the fundraising their PTO does. And in all the conversations, it went without saying that parent support was a benefit to their students, who rarely had the out-of-school challenges or academic deficits of those in district B.

District B staffs' discussion of parents was nearly the polar opposite of that in district A. There was no discussion of parent pressure to increase performance. And opt-

outs were a non-issue. One principal put it bluntly: “We had zero opt outs. No parents even asked. [The] community trusts administrators. [It’s a] different culture. [It] doesn’t occur to them to question it.” A teacher echoed this but for parents’ trust of teachers: “Parents will believe anything I say about their kids. I was scared my first year. I told my husband, whatever I say about their child, they will believe. They trust me. That is a huge responsibility. Telling them their kid is failing will matter. Telling them they will go to college matters.” Another teacher was quick to correct that the general sentiment about parent involvement did not mean parents didn’t care. Instead, they thought it was a product of lack of capacity to challenge it. “They just don’t know how to.” While this may limit the pressure parents exert on administrators and teachers however, the corollary is a clear detriment. As one teacher stated when addressing her students’ needs, parents aren’t able to support their students academically: “parents don't read to kids, don't talk to kids... Our kids’ parents don’t even read or write.”

Relationships between Teachers and Administrators

Given the role of administration in carrying out the requirements of New Jersey’s reforms, the relationship between teachers and administrators plays a pivotal role in how staff experience and respond to reforms. It is also one of the major differences that emerged between district A and B. In district A, teachers generally have positive relationships with principals and trust their competence and intentions. Likewise, teachers and principals generally feel good about the central administration. To the extent that they dislike aspects of reform, they blame the state, not the leaders of their district. In stark contrast, district B teachers have more complicated relationships with their principals and an almost universal lack of faith in their central administration, trusting

neither their competence nor their intentions. While they recognize the state's role in pushing reform, it does not let their leadership off the hook. In a negatively reinforcing loop, reforms amplify their animosity towards their leadership while their distrust of leadership amplifies their frustration with and lack of confidence in reforms. Having been a participant observer as an administrator in district C, this difference is not unique to district B. Distrust of administration dominated many interactions there as well.

When speaking with teachers and principals in district A, it is clear that they trust and respect each other and generally have favorable views of their central administration. The sense is of a well-integrated organization of staff with different roles working towards a common goal. Central administrators see their role as creating the conditions for school success, including supporting staff and protecting them from outside influences that might interfere with their work. Principals likewise seem to view themselves as supporters of their teachers, from coaching to shielding them from undue outside pressure. The result is that staff feel supported and respected with the resources to do their jobs. I address resources in the next section. In this section I focus on how this insulates them from the most negative side effects of reform.

While district A staff did not have positive views of New Jersey's reforms, the positive relationships between people at all levels of the organization contributed significantly to ensuring they viewed it as a nuisance, not an existential threat. One principal summarized why reforms did not create tension between school staff and central administrators: "I never look at it as coming from [the Superintendent]. This is coming from the state/county." Others echoed the sentiment; they blame the state for its interference, not their administration. This perhaps reflects a concerted effort by central

administrators to make staff feel protected from outside interference. Several teachers noted how the administration's message was not to worry about PARCC, for example. Said one teacher, "the whole district is taking a wait and see approach and telling teachers not to worry." Another noted they were generally assured nothing would trigger automatic actions like firings. While this is technically not true - as discussed above, the law requires tenure charges under certain conditions - as a practical matter the statement is defensible because, as one teachers stated, "no one here is below effective." This also highlights the interaction between organizational dynamics and student needs. High performing students provide the organization the luxury of not being overly concerned with accountability metrics. The administration, in return, takes advantage of that luxury to protect its staff.

This goes well beyond central administration as both principals and teachers talked about the role of principals and supervisors as supporters and protectors of teachers. One principal talked about how she uses social media to promote her teachers because "I really believe in what they are doing." Another described her role as working together with teachers to get through challenges: "I was in it with them. We spent a faculty [meeting] and I'd come and sit with them for 40 minutes. Let's do this. If you tweak this, let's do this." She said one of the key benefits is that teachers feel less threatened. "I know she is not gonna kill me if I don't this right." She also acknowledged the role of competence in making teachers feel comfortable. "I'm organized. I don't know that all the other buildings feel that way if their administrator is not as organized. They might feel a little more frustration. I do know some of my people that have different supervisors that they report to might be a little more frustrated." She then acknowledged

that it is fair to say the new requirements, particularly regarding documentation, put a high priority on very well organized administration and increase pressure on administrators to be more organized.

For their part, teachers corroborated principals' account of their own supportiveness and competence. Said one teacher, "My principal and supervisor are very good about helping me tell the story." Said another, "[Our] supervisor told us not to stress because PARCC is so new." One teacher was effusive in praise for her principal. "[My] current principal relationship is good. Her expectations are high and clear. My meeting with her last year was good. I am very, very, very supported by her. She totally gets where I'm coming from." Reflecting on the competence of administration as an asset, one teacher noted the luxury of not having to worry about curriculum: "We have a great math program. The standards piece I don't have to touch. The curriculum is great and aligned." And finally, a special education teacher who is also a union representative reflected positively on the development of the teacher evaluation tool, a key aspect of reform:

> [The administration and staff had a] very good relationship in the creation of the DEAC [District Evaluation Advisory Committee]. [We] had to originally come up with [an] evaluation tool. We went with Danielson. which mostly closely aligned with what we were already doing. [It] makes for easier transition to [the] state mandate.

This is not to say that they loved all aspects of implementation. The union representative acknowledged that the FFT rubric is not ideal for summative evaluation, for example. They key though is that staff do not resent their administration as a result.

One teacher summarized how all of this plays into how they feel about working in district A:

> [I] love it here. [I feel] very fortunate. This is a great school district. [It is] incredibly supportive of teachers. [They] try really hard to provide everything

> they can. [I] enjoy the administrators, the building principals. [I have a] great rapport with both [administrators]. Staff here is wonderful.

This combined with a comfortable and safe setting, high-performing students, involved and supportive parents and, as discussed below, resources that make work easier while shielding them from state interference, paint a picture of how district A staff may avoid the worst consequences of reform.

The same cannot be said of district B. Whereas the relationship between staff and administrators in district A is an asset, the relationship between staff and administrators in district B is the opposite. Rather than feeling supported, staff feel overburdened. District A staff rely on well-developed curricula that they trust. District B staff create their own from scratch. Rather than feeling protected by administration, staff feel threatened. Teachers trust neither the competence nor motives of those to whom they report, a toxic setting that is exacerbated by reforms that mandate evaluations that require skilled and well-intentioned evaluators and increase the importance of the resulting evaluations. Teachers in district B do not feel that their central administration is willing or able to put them in a position to succeed and often do not think their principals are going to be good evaluators. Both contribute to their sense that the system holding them accountable is fundamentally invalid and unfair.

One of the most telling differences is how little principals in district B had to say about supporting their teachers or working with teachers or their central administrator. As in other areas, this may reflect a difference in the selection of principals and the perceived role of principals in districts like district B compared to district A. Instead of supporting teachers, for example, the role of principals in district B seems more about sorting: identifying good and bad teachers and acting accordingly. While this is far more

consistent with the implied theory of action of accountability reforms, as discussed in chapter 6, it may counterproductively limit the effectiveness of reform.

In contrast to their principals, teachers had a lot to say about administration. One teacher highlighted just how differently the central administration is in district B than A, leveraging the threat of reform rather than shielding them from it:

> Stop with the state takeover. Its held over our heads here. If we don't get our scores up, [the] state will take us over. Our admins say this. We have to document ...to prevent state takeover when state looks. This is why we have to do stupid three part objectives. Because our children are not on grade level. They are not going to catch up without assistance. [The t]eacher is stuck.

Another teacher described the importance of having competent administrators for evaluation to work as intended. "If your admin is not properly trained and doesn't know what they are looking for, it doesn't matter what the rubric [is]." That teacher noted that she is lucky because now she has a good principal and VP and therefore good evaluations. Her counterpart gave an example of her lack of faith in the administrations' competence, an example that stands in stark contrast to the asset of a strong curriculum identified by district A staff:

> The district's math book was terrible. It was not aligned. It was horrible. I trusted it okay before I was in tested grade, but once got into a tested grade, I said, I'm not trusting this. I'm not putting my name on this. We decided we would teach for mastery and built our own curriculum, ignoring the district.

Another teacher addressed feeling unsupported: "Math and ELA specialists [are] supposed to help but [they are] not doing [it] at this school."

Two other statements summarize the overall feeling about administration, seeing them as more of a threat than a support: "A lot of people fear administration. They don't want to be the one to take that first step. It's a risk your taking. There's a target on you." Another drove home the perception that teachers would be better off going it alone than

having the administration they have. “They don’t care how good we are. We can’t change anything. [The] principal is validated because of me. You don’t need her. You need 100 of me. They aren’t using the standardization right. I am the strong. Why aren’t you using me?!” Others acknowledged that this teacher was high performing. She was extremely devoted and used data in a way that is entirely consistent with the vision of reform. She also confided in me that she was planning to leave the district to go to a suburban school district, a consequence I address in both chapter 6. This highlights a key point. Whether or not their perceptions of administrative competence or motives reflect reality, they have real implications for outcomes and should therefore be accounted for in any effort to change behavior.

Resources

A common thread connecting both the fact that teachers feel more supported and satisfied in district A and less threatened by state intervention - discussed further below - is access to sufficient resources. Greater access to resources that make it easier for teachers to do their jobs reduce the sense that accountability will unfairly punish them for factors outside their control. While the threat of state intervention is highly linked to local financial resources, the resources identified by staff as supporting them in their work are not necessarily financial. Rather, many key resources are intangibles that can be leveraged to improve the quality of staff and materials independent of financial cost. One of the most influential examples of this is the desirability of working in the district. This can both increase the supply of teachers who wish to work there and decrease the rate at which staff leave, both of which allow the district to be more selective. A related resource is administrative efficiency, which further contributes to selectivity by, for example,

allowing district A to recruit staff substantially earlier in the year than district B. Another resource is organizational culture, which impacts everything from expectations to the degree with which staff are willing to help each other. I therefore define resources here broadly to include all assets that make the work of educating students easier. Financial resources play a part but it is not as dominant as might be expected.

Several staff members in district A implied that being in such a desirable district allowed district A to select only the best staff. One teacher put it directly when discussing why she was not concerned with evaluation scores: "I wouldn't be working here if I wasn't getting those [high] scores. The expectations are high here in the district. My feeling is people are here because they are good and will get good numbers anyway." Another teacher implied something similar and added the role of internal supports in helping ensure that the staff get good scores.

> We are high achieving type A people district. People believe they should get 4s over 3s. We did not have any ineffective or partially effective staff last year. [I] wouldn't expect [us] to. We go above and beyond all the time. We also have a lot of internal structures to support each other. We reach out to each other. [We] share ideas. [There is] lots of peer collaboration.

Internal supports include the quality of the materials that the district makes available. Here again, district A staff felt well supported, noting the quality of the curricula and how their alignment to standards takes one burden off their shoulders.

Another resource is in some ways the flipside of students with fewer needs. District A and district B are not fundamentally that far apart in annual per pupil spending, with both well over $20,000 annually (Taxpayers Guide 2019). However, district A can allocate those resources to things like smaller class sizes and extracurriculars that enhance both student and staff experiences. One principal noted that very few teachers in

her building even had mSGPs because their classes were well below 20 students. A teacher reflected on how much she enjoyed teaching orchestra. In addition to impacting how selective district A can be when hiring, resources may also limit turnover and keep demand for new staff low. A principal identified at least one significant benefit for her in this. She had no first year teachers, reducing the burden on her to on board and support inexperienced staff and reducing the number of observations she had to do.

In contrast, her counterpart in district B lamented that nearly a quarter (eight) of her thirty-five teachers were new, which made her observation burden feel unsustainable. Another principal described for me that most of their hiring does not occur until August, after the majority of the most qualified candidates have already found jobs. At the same time, teachers noted the sheer number of students they had along with their needs (as discussed previously). While I did not sample randomly, I note here that the average class size for the teachers I interviewed in district B was nearly 26, compared to under 19 for district A. Again, this is not necessarily a product of lower per pupil spending but instead competing demands for resources because students have greater needs. I did not capture evidence to validate whether it reflects allocation decisions of a less competent administration, as teachers' frustration with administration might suggest. Whatever the reason, staff seemed to experience the constraints, with one teacher noting that her school's funding was recently cut. Finally, whereas district A staff spoke highly of their curriculum, district B staff complained of not trusting their materials and creating their own, as discussed above. Teachers in district B thus feel substantially less supported while trying to serve more students with substantially higher needs. In this light, it is not

surprising that they feel more burdened and threatened by the state's accountability reforms.

Relationship between District and State

The final area that negatively impacts district B staffs' experience of accountability reform is the relationship to the DOE. It is here that financial resources play perhaps the largest role. While total per pupil spending is not vastly different between districts A and B, the source of funds is. District A gets the vast majority of its funding from property taxes. District B, on the other hand, gets the majority of its funding from the state. As a consequence, district A feels very little pressure from the state: "Pressure comes from parents, not the state." District B staff, on the other hand, seem to internalize a perpetual threat of stronger state intervention. As one principal said, "In general, yes there is more pressure from the state, they could take the money away. If continue to fail to show growth, we could lose Title 1 or SPED money." This pressure clearly made its way to teachers. One teacher described the tangible impacts beyond just feeling pressure:

> My sense is that in a district like this you are very beholden to the state. [There are] so many rules and regs [that you] lose some momentum. [We are] so much more vulnerable here. [We are] slaves to it. [We] spend a great deal of time documenting compliance so that when the auditors come, we can show we've done everything...

As noted, the teachers perceive the administration as amplifying rather than mitigating the state's influence, further exacerbating the difference in the experience of reform between district A and B.

One district A teacher inadvertently captured the core result of all the differences in context when she said, "all the pieces fit together for schools like ours. [That's] not likely [true] in a place like [a low-income urban district]."

4.4 Conclusion

New Jersey's education reforms thus meet all three conditions of a failed effort to make a social process legible to government. To facilitate management through accountability, the reforms substantially simplify a highly complex social process. This is reflected in many of the issues principals and teachers in both districts have with the evaluative aspects of reform, including a strong perception of invalidity. Achieving standardization and comparability also led to centralizing expertise, shifting it further from those situated in the schools and classrooms and closer to state bureaucrats, technical experts and national organizations. This enhances perceptions of invalidity in both districts and also tends to make staff feel undervalued.

While staff in both districts experience the simplifying and expertise-centralizing aspects of reform, the degree to which it impacts their experience of their jobs is very different. Owing largely to dramatic differences in context, staff in district B feel far more threatened by accountability measures that do not take into account the fundamental differences in the challenges they face relative to districts like district A. With the conditions of failure satisfied, the next question is whether those conditions are triggering the types of responses that we would expect to result in the ineffectiveness researchers have identified.

5 Empirical Support: The Conditions of Failure Trigger Mechanisms of Failure

That New Jersey's education reforms so thoroughly meet the three conditions of failed legibility efforts suggests is only the first step in establishing this framework's utility. For the framework to be useful, it must also explain exactly how these conditions contribute to the failure of education reforms to deliver promised results. Derived from the literature summarized in chapter 2, Table 5 summarizes a proposed relationship between the conditions of failure and the mechanisms by which those conditions might lead to failure. I group the potential mechanisms by which the three conditions may lead to failure into four categories.

Table 5.0 How the conditions of failure trigger the mechanisms of failure

Condition	**Mechanism of Failure to Achieve Goals**
Oversimplification	● Mischaracterization of the problem ● Engendering Resistance
Devaluation of Local, Situated Knowledge	● Mischaracterization of the problem ● Engendering Resistance ● Undermine the Conditions of Success ● Poor Implementation ● Turnover
Ignorance of Context	● Engendering Resistance ● Poor Implementation ● Turnover

The relationships are not one to one. For example, both oversimplification and devaluation of local, situated knowledge can lead to mischaracterization of the problem. Likewise, all three conditions of failure might engender resistance in key stakeholders. The framework I am testing does not require a simple, linear relationship from each

condition of failure to actual failure. Rather, the framework suggests that the combined conditions of failure collectively trigger enough mechanisms that get in the way of success to explain why reform has been so disappointing to so many.

5.1 Reforms mischaracterize the problem

I begin with mischaracterization of the problem because it is the most straightforward source of failure: a solution that does not address the root cause of the problem has little hope of solving it. The connection between this mechanism and the conditions of failure is also straightforward. Where those trying to solve a problem oversimplify it and/or ignore local, situated knowledge, they are very likely to mischaracterize the true problem. This was the case in Mitchell's (2002) example of efforts to solve the supposed food shortage in Egypt. Relying on oversimplified tropes that local stakeholders could easily have corrected, international aid organizations mischaracterized the cause of the problem as a simple geographic limitation and, as a result, failed to solve it.

As noted in chapter 4, NJ's education reforms bear strong analogy to the cases in Mitchell (2002). The goal statement in the TeachNJ Act implies reliance on the simplified trope that has come to dominate the education debate: educational failure is the result of teachers not being effective enough and they are ineffective in part because their performance is not measured well enough. Or as the legislature phrased it, they believe they can "raise student achievement by improving instruction *through the adoption of evaluations*..." (NJAC 18A:6-118:2.a emphasis added).[28] Assuming for the moment that

[28] The full preamble reads: "2. The Legislature finds and declares that:

better instruction is enough to improve achievement, for the theory to be correct, evaluations must also lead to improved instruction. For this, the legislature offers several mechanisms: evaluations will provide useful feedback to educators, improve professional development and lead to better decisions on whom to retain and let go. (NJAC 18A:6-118:2.a). Each of these has its own implied causal assumption: professional development and feedback assume teachers lack the knowledge to be effective while improved personnel decisions implies that the composition of the workforce is the problem (i.e. that there are some teachers that cannot become good teachers through professional development and targeted feedback). These are all to differing degrees consistent with what I learned from educators in District's A and B. Educators with local, situated knowledge tend to agree with the argument that at least part of the problem is knowing what to do. If the legislation were consistent with its preamble, that might be the end of this section. It is not.

The legislation and implementing AchieveNJ regulations, both as written and as implemented, are dominated by a different, implied causal assumption: that the real problem is that educators lack sufficient incentive to be effective. Legislation and regulations consistent with the declarations in NJAC 18A:6-118:2 could rely predominantly on formative evaluations to provide the information that educators and

---

a. The goal of this legislation is to raise student achievement by improving instruction through the adoption of evaluations that provide specific feedback to educators, inform the provision of aligned professional development, and inform personnel decisions;
b. The New Jersey Supreme Court has found that a multitude of factors play a vital role in the quality of a child's education, including effectiveness in teaching methods and evaluations. Changing the current evaluation system to focus on improved student outcomes, including objective measures of student growth, is critical to improving teacher effectiveness, raising student achievement, and meeting the objectives of the federal 'No Child Left Behind Act of 2001.'"

decision-makers are lacking. Instead, the legislation and regulations are dominated in both theory and practice by highly prescriptive summative accountability measures. As detailed in chapters 1 and 4, TeachNJ and AchieveNJ require that teachers receive a summative score and that in key circumstances defined by those scores, discretion must be removed from personnel decisions. Moreover, there is little in the legislation requiring that the information derived from evaluations be used for anything other than summative purposes. For example, to be used formatively, test results would need to be available in reasonably short timeframes and the reports would need to contain highly detailed student-level data. The legislation has no requirement that either of these occur (as discussed in section 6.1.1.2, educators in both districts repeatedly expressed frustration about their inability to use the testing data to improve instruction).

Thus, despite the legislation's characterization of the problem as at least partly a lack of knowledge or information, the solution crafted implies policymakers believe the problem is a lack of incentive. At the same time, the nature of accountability as implemented implies that educators can be incentivized by retention decisions, i.e. that they are motivated by self-preservation and/or career advancement. The question here is whether either of those is a mischaracterization. Are teachers less effective than they should be because they don't have the incentive to be more effective? And if so, can they be motivated by self-preservation? Both the literature and what I learned from educators in districts A and B suggest that most teachers have sufficient incentive to succeed because they are motivated more by a desire to help students than by self-preservation. NJ's education reforms therefore likely mischaracterize the problem.

This is another area in which districts A and B experience reform differently. The degree to which this mischaracterization portends failure differs between the two districts. This is a byproduct of ignoring context. In District A where teachers are better insulated from accountability measures by district policy and the greater likelihood that their students will succeed, they are impacted by accountability only to the extent that they have an innate desire to have a rating that reflects their perceived value. In District B, there is a far greater perception that summative accountability is an existential threat.

5.1.1 Teachers are generally sufficiently motivated by the desire to help students

The literature suggests that educators, like other public servants, are generally motivated by a desire to benefit others, rather than to directly benefit themselves. Ingersoll (2003), for example, notes "the unusual degree of commitment of those who enter the profession." (p. 236). Teachers in general are more motivated by non-monetary rewards than other workers. (Ingersoll 2003). As with other public servants (Lipsky 1980), teachers have a high public-service orientation and place a great deal of value on intrinsic rewards and the sense that they are making a difference. (Lortie 1969, Lortie 1975, Ingersoll 2003). Because they care so much about their students' success, rather than being under-incentivized, many teachers may be stressed by the burden of trying to help their students succeed. Lortie (1975) notes that they tend to subject themselves to demands that are higher than are achievable.

Interviews with District A and B educators do nothing to contradict this literature. Teachers in both districts A and B repeatedly expressed that their reason for teaching is to help students and neither teachers nor principals suggested that staff lack motivation to do well. As one district A teacher noted, "that's why people stomach it. Because of the

kids." Similarly, a district B principal reflected the general sentiment that "this particular staff has always wanted to do well."[29] And there is concrete evidence that as a result of being motivated by other factors, they do not attend to their ratings in the way reformers might have envisioned. Echoing Lortie's finding that teachers are far more likely to value their perceptions of success than those of central authorities, comments like this one, from a teacher in district A, were common: "I don't focus on numbers. Regardless of numbers I'm going to do what I think is best." This comment from a district B teacher expresses the same sentiment: "I don't know my score yet but to be honest, I could care less [about my score]. I know what's going on in here." Others said they "push teacher evaluation to the side" or "ignore ratings for the most part." In fact, many teachers and principals couldn't be bothered to learn the details of the evaluation system. While this was common in both districts, it was more prevalent in District A, where they are more institutionally insulated.

5.1.2 Demand for formative support suggests the larger problem is that educators lack the information they need to succeed

If educator motivation is not the problem, what might be? Educators did not directly address this question and a detailed consideration of the issue is beyond the scope of this dissertation. For this framework, it is sufficient to establish that reform's underlying assumptions about incentive are wrong. Educators did, however, volunteer

[29] This principal did, however, note that evaluations helped highlight which teachers were not performing well.

strong preferences for formative tools. I therefore infer that the primary problem is that educators do not know how to succeed or do not have sufficient resources to do so. [30]

Educators demonstrated strong preferences for resources that help guide instruction. This preference manifests in at least three forms: frustration that summative requirements inhibit or preclude formative support, the generally positive view of CCSSs and the support for FFT when used only for its intended, formative purpose. When probed about each of these, principals and teachers responses indicated that what is really lacking is a clear guide on how to help students.

There are two common examples of frustration about summative requirements interfering with formative development, consistent with the Stecher et. al.'s (2018) findings. First, many principals expressed frustration with the fact that summative observation requirements limited their ability to do the informal, formative observations they think are critical. One principal's explanation highlights that there are two ways in which summative observation requirements get in the way of formative support: time and a shift in how teachers respond. The principal explained that there is "no time for informal walkthroughs because formal [walkthroughs] are so time consuming. This year I have 70 formal walkthroughs." The summative nature of the observations adds time because it requires pre and post conferences and entering structured data into a digital tool. It also sparks something that formative observations might not: demands from teachers to challenge results.

The second area in which summative requirements interfere with formative support is standardized testing. To accommodate the use of PARCC as a summative

[30] They were not referring to financial resources or materials, though some district B teachers did talk about using their own money for materials. Rather, the resources were more like administrator support, stability, curricula and other tools to guide instruction.

instrument, state contractors score the test and the results are not available until the following school year. When released, the results are not provided in a way that gives teachers information they need to customize their instruction based on their students' performance. Teachers and principals in both districts expressed frustration with this. As noted in chapter 4, they relied on additional evaluations for formative purposes, but that added testing time. As one District A principal noted, "testing time is ridiculous," estimating it at 45% of the time and 60% of school days.

There are two things that educators liked, however, that further imply a preference for formative support over summative accountability. Contrary to a common narrative, teachers and principals in both districts A and B nearly universally appreciated CCSSs. Because this was so counter to the public narrative, I probed to learn what they liked about them. Simply put, CCSSs were welcomed because they informed instruction. Teachers liked that CCSSs gave them a roadmap, making explicit what they needed students to understand by when. They also liked that it facilitated alignment across grade levels. Given how contrary this finding is to the public narrative, I summarize my notes about educators' feelings about CCSSs in table 5.1. Note that several educators explicitly rebutted the political narratives. One takeaway is that those promoting the narrative that CCSSs are unwelcomed by educators are as guilty of ignoring local, situated knowledge as the advocates of reform. The risk is that one of the few welcomed elements of reform may be withdrawn or changed to score points with a public audience that has little information of the real benefits CCSSs provide.

Table 5.1 Teachers' and principals' statements in support of common core

I think its ridiculous to say it takes away discretion. Any standard is general enough to let you work in your activities, lessons, experiences etc.
I don't dislike common core
I like common core. I like the alignment. No problems going from one place to another. Makes it easier. Has not limited my freedom. I'm still capable of doing projects etc.
Common core lets us share ideas across classrooms and limits reinventing the wheel. Can now do projects across the district
No problem with common core
I think the CCSSs are good
CCSSs give newer teachers a better rubric to know what they should be doing. Provides a guided commonality. That part affects teaching. I for the most part agree with the CCSS because can't do one shoe fits everyone. Provides a focus. Provides an understanding. I think teachers would use CCSS as a guider of skills even if it weren't tested. CCSS is meeting a demand.
Common core politicized but teachers like it
I like common core. Focused. Structured. Especially if its PARCC aligned
Testing sucks but the standards are good. Lots of prep time but that will decrease as get more experienced
I don't see an issue with Common Core. Cause and effect are important things. Skills that every student should know. The way you teach it is up to you but the baseline is important.
I need something to go by. If they took away the CC, the SGOs, the obs etc. I don't think I could create my own guidelines. I like knowing what my standards are. Then hitting them how I want, subject to curriculum.
I feel like Common Core actually helps me do the things I want to do. It has stuff about college ready and career ready. Its showing people how important it is. How they choose to do it.
Common Core fine. Just don't like the way it's evaluated.
Curriculum is so much more structured. Good in that you know what you need to do.
CC is good and beneficial but not the be all and end all of education.
I'm not a fan of PARCC but I'm a fan of the common core. It's very concrete.
Common Core and PARCC shouldn't be grouped as same thing. PARCC sucks. I have no problem with Common Core. I like that its focused and I know exactly what to teach. Now I know I have to teach 44 things in 45 weeks. It really simplifies planning.

The take on FFT was a little different but equally reflects a preference for formative resources that support instruction. As detailed in chapter 4, when teachers and principals had issues with FFT, it was only as a tool for summative evaluation. Their challenges were largely about its validity for that purpose. They generally found it to be a valid and helpful guide to high quality instruction. As one district A teacher said, "Danielson is good formatively, not summatively. It provides helpful structure to evaluations." Her colleague agreed, "I think it does reflect best practice." One teacher in district B noted that FFT had helped them get better at formative assessment itself.

Thus, while New Jersey's reform law and its implementing regulation primarily target incentives, the evidence from districts A and B suggests incentive is not the problem. To the extent that it is, educators are not so motivated by self-preservation that accountability measures are likely to work. Instead, preferences for formative measures suggest that the problem is much more likely to be that educators lack the knowledge, tools and conditions to succeed. The fact that the push for legibility resulted in reformers focusing on the wrong source of the problem may partially explain reforms' lack of success. The evidence suggests that a better approach would be anchored in formative, rather than summative, measures, consistent with the recommendations of Rowan and Raudenbush (2016).

5.2 Reforms are engendering resistance

Oversimplification, devaluation of local expertise and ignorance of context might also get in the way of success by engendering resistance. If that is the case, there should be evidence of resistance in districts A and B. There is. I hypothesized that accountability measures that limited educators' discretion, and especially those that impacted their

ability to define success would engender resistance. I further hypothesized that resistance might include loose coupling, turnover and unproductive coping mechanisms. Finally, I hypothesized that the degree and nature of resistance would differ based on context, with loose coupling better preserved in districts like A and turnover and unproductive coping mechanisms a greater risk in districts like B and C. I address the evidence regarding turnover in section 6.1.5. Here I present the evidence of resistance and the degree to which it differs in different contexts.

Resistance in district A manifests itself as a preservation of loose coupling. At the principal and district level, administrators went out of their way to shield teachers from feeling overly burdened by accountability measures. District support was captured in statements from district A teachers like "I never look at it as coming from [the Superintendent]. This is coming from the state/county" and "the whole district is taking a wait and see approach and telling teachers not to worry." The protection was tighter for principals, with common statements like "My principal and supervisor are very good about helping me tell the story" and "[Our] supervisor told us not to stress…" discussed in detail in section 4.3.2. This shielding enabled teachers to resist in fairly passive ways, most notably by "pushing teacher evaluation to the side" and "doing what I think is best" regardless of the ratings and "not changing anything." The freedom to respond minimally to reforms afforded by loose-coupling is a luxury of being in a district with high-performing students and financial independence from the state. It is thus not surprising that I saw little evidence of the same protections in district B. Rather, no district B principal spoke of protecting teachers and teachers spoke of administrators holding a potential state takeover over teachers' heads as a threat to get scores up.

That said, not all forms of resistance in district A were passive and tied to loose coupling. District A teachers also engaged in what one principal called "subtle gamesmanship." She was quick to distinguish this gamesmanship from "cheating" such as what took place in Atlanta. Rather, she referred to "teaching to the test" and "coaching" students in advance of the test. For their part, teachers added candid admissions that they do game SGOs: "We chose SGOs we knew our kids could perform well on. Also, you can just teach the stuff again the day before the post-test." The teacher who said that did note that the principal can and has revised SGOs but the sense from my interviews was that SGOs were little more than an administrative box checking exercise. One teacher in district A also spoke of gaming the observations: "Other teachers just choose a student-led lesson when [they] know an observation is coming."

While these forms of resistance were also present in district B, what distinguishes it from district A was resistance to district policy, not just state reforms. Examples ranged from pushing against mandatory pacing ("We push it. If they give us three days to teach that skill, sometimes we'll take 4, we'll just short another skill.") to ignoring mandatory grading rubrics ("District policy is [that in-class assessments] are supposed to be considered major assessments…worth forty percent of [a student's] grade. But my kids do horribly. So I don't enter it as a major assessment. Only as classwork, which is thirty percent and gets averaged in with other things.").

In addition, while there was no evidence in district B of principals shielding teachers, I did witness a form of principal resistance in district C that, though it may have been done more as a way to cope with the burden of observations than with the intent of shielding teachers, could have had a similar effect. While doing routine reviews of

observation scores in district C, we noticed that several principals had a distinct pattern in their observations. All of their teachers received threes on all four FFT domains and all 22 items. Several other principals had only minor variations. Overall, over 1/3 of practice scores fell within .1 points of 3 while the overwhelming majority of scores fell within the "Effective" range (2.65-3.5). Table 5.2 summarizes District C's teacher practice scores from 2013-14 through 2016-17. At the same time, the correlation between practice scores and mSGP was .26. While it is possible that the practices measured by FFT do not lead to greater success on PARCC – something that would be problematic for one or the other metric – the concentration of scores close to three suggests it is more likely that principals were defaulting to "neutral" scores. It is difficult to infer the intent behind such a high concentration of effective ratings, but the effect was to make it impossible to use observations – the largest share of a teacher's evaluation – to distinguish teachers. During the four-year period, the correlation between practice scores and summative scores was .9.

Table 5.2 Share of district C teachers with practice scores near 3

2013-14 to 2016-17	**Number**	**% of Total**
Total Annual Teacher Practice Scores	3951	100
Rated Exactly 3	87	2.2%
Between 2.99 and 3.01	214	5.4%
Between 2.98 and 3.02	365	9.2%
Between 2.9 and 3.1	1357	34.4%
Effective (2.65-3.49)	3232	82%

Further, while a detailed analysis of statewide microdata was beyond the scope of this dissertation, NJ DOE's initial progress report suggests that, statewide, the evaluation system was not being used to sort teachers as intended. In 2013-14, over 97% of teachers were at least effective and nearly three quarters of all NJ teachers were rated exactly effective. In 2014-15, those numbers were roughly two-thirds and over 98% respectively.[31] While the state report characterizes this as evidence that all teachers are performing well and even improving, my experience along with the large role observation scores play in summative ratings suggests it is at least as likely that observers' tendency towards observation ratings of three explain the results.[32] Notably, despite arguing that teacher practice is improving and having a theory of action that teacher improvement would leave to outcome gains, there is no mention of outcomes in New Jersey's report.

There was thus resistance in all three districts, but it took on different characteristics. There was one form of resistance, however, for which I found no evidence. The technical discussion of SGPs includes a note about the fact that they can be gamed: "Like gain-based models and, more directly, residual gain models, SGPs can be artificially increased by deflating initial year scores. In the intuition of SGPs, this deflation changes the academic peer group of students to one that will tend to be lower scoring, resulting in an inflated SGP. As a corollary, this will also influence percentile growth trajectories." (Castallano, K. and Ho, A. 2017). I found no evidence that teachers or districts engaged in this kind of gaming. Based on what I learned, my hypothesis is

[31] The Implementation Report can be found here: https://www.state.nj.us/education/AchieveNJ/resources/201415AchieveNJImplementationReport.pdf

[32] The report does not break out observation scores in its Key Findings.

that there is neither the technical understanding of the models nor the alignment of incentives between district administrators and/or teachers across years to make this a viable strategy.

5.3 Reforms may be undermining conditions of success

So far it appears that accountability based reforms in education may fail to achieve the goals of their proponents because they mischaracterize the problem and engender resistance. However, even if the problem were correctly diagnosed and the solution designed and delivered to engender buy-in rather than resistance, that would not be enough to portend success. As highlighted in section 2.2.1, accountability measures aimed at street-level bureaucrats, including teachers, frequently have the unintended consequence of undermining the conditions these public servants need to be successful. The literature suggests four conditions that are key to the success of teachers are particularly at risk from accountability: control and flexibility, legitimacy and authority, collaboration amongst teachers and motivation. I found substantial evidence that reforms are negatively impacting all four conditions. In addition, I found evidence of two additional responses to reform that seem inconsistent with the goal of improving student outcomes. First, summative accountability measures engendered skepticism of data that undermined the ability to use data formatively. This parallels the tension Stecher et. al. (2018) noted in attempts to use evaluations both summatively and formatively. Second, testing seems to be causing unproductive and potentially unhealthy stress in students and creating a transactional environment that might not be conducive to genuine learning.

5.3.1 Control and Flexibility

As noted in 2.2.1, the highly variable, context-dependent and judgment-intensive nature of teaching (Sizer 1984) may require teachers to retain control over their classrooms with the flexibility to adjust as needed. Ingersoll (2003) suggests that poorly designed accountability systems might unproductively interfere with both control and flexibility. I found the evidence for this mixed, with many educators lamenting that prescriptive rules sapped needed freedom while others indicated that they retained the necessary discretion to do their jobs well. There was, however, near universal agreement across districts, teachers and principals that reforms imposed administrative burdens that led to real limits on educators' ability to serve students. This may be one of the larger impacts of a failure to heed local expertise and context. As with so many other consequences of reform, here too there seem to be noteworthy differences between districts A and B, with district B teachers more likely to feel constrained. Conversely, district A principals expressed more concern than did district B principals.

Educators identified several mechanisms by which New Jersey's reforms may be reducing needed control and flexibility. Three of four district A principals noted two mechanisms in particular: enhanced pressure towards standardization, especially across schools, and teaching to the test. The latter appeared to be of secondary concern. The first was highly contentious. In the interest of respecting the request of an interviewee to "be careful how you print this," I will summarize discussion of district A's efforts to align across schools at a higher level than my interviewees did. The gist of their statements is that individual schools have had to scrap valued programs that distinguished them from each other in response to pressure arising from parents' review of state testing data.

Differences in aggregate test results across schools triggered demands for changes to which the district appears to have at least partially acceded. One principal who acknowledged the value of alignment more generally lamented, "my biggest frustration as an administrator is the constraints that have been placed upon me and my building for me to carry out my vision. Should [a high quality education] look the same from building to building? "No." Other principals echoed this and added that "there is a sense that teachers feel a loss of freedom from the new aligned curriculum," that teachers respond to pressure that comes from differences in test scores across schools, that you "can't do anything that can't be standardized," and notably that it is "affecting teacher practice the wrong way." Here staff distinguished curriculum alignment from the standards-alignment driven by CCSSs. Their concerns were about pressure to standardize curriculum – the means by which the standards were taught – driven by standardized testing, not CCSSs.

For their part, district A teachers did not seem to corroborate principals' perceptions of their limitations. While they acknowledged a loss of discretion and expressed some concern that prioritization of math and ELA took away from other subjects, on balance district A teachers seemed to feel like they retained sufficient discretion to do what they needed to do. The general sentiment was that while CCSSs, PARCC and their evaluations impacted *what* they taught, they were not limited in *how* they taught it. This may be a byproduct of district A principals protecting teachers. For example, one teacher noted that the administration "really encouraged... project-based learning."

In district B the roles were reversed. Teachers were far more concerned about the degree to which reforms penetrated their classroom practices than their principals. Not a

single principal in district B expressed concern about lost discretion. The only related comment was that CCSSs are more likely to enhance creativity than to restrict it, echoing district A teachers' sentiment that CCSSs focus you on the what, freeing you to be creative about the how. Conversely, several district B teachers expressed a sense of a far more impactful intrusion. One reflected on "the inner-attacks on the classroom," describing the prescriptive nature of how district B was implementing reform. Another lamented, "they are trying to force you into more of a mold and not allowing you as much freedom." Meanwhile, a teacher from one of the higher performing district B schools echoed the concerns of district A principals, noting that her school had to change a well-liked curriculum to ensure alignment with the rest of district B's schools. Their concerns were also more likely to focus on accountability than on CCSSs. Like their district A colleagues, district B teachers expressed that CCSSs did not hinder their creativity. "The way you teach is up to you. But there is so much freedom with the common core and can incorporate the lessons in my own way."

While agreeing that CCSSs do not limit creativity, all interviewees also agreed that the administrative burdens of New Jersey's reforms were highly problematic. Their responses suggest that time constraints related to these administrative burdens likely impacted educators' control and flexibility more than any substantive changes did. In the absence of accountability, however, the administrative burdens would likely have been easier to ignore and less problematic. For example, one district A teacher noted that she spends an inordinate amount of time on artifacts because "if it's not on paper, [they] can't score it." A full digest of all comments related to the time taken to comply with Achieve NJ's requirements would make this dissertation unreadably long. Instead, I summarize

the types of time burdens educators noted along with a few representative examples, in table 5.3. Perhaps the biggest impact is on job satisfaction, as the administrative burdens were among the top complaints of teachers expressing a possibility of leaving their district or the profession. One district A teacher summed up the overall burden, perhaps hyperbolically, though I did not get the sense her colleagues would disagree. “In a five year period, overall scale is more than 10 times what we've done in the past. At least that much more time.” The implication was likewise common: “I take things home.”

Ultimately, while educators were broadly comfortable with the freedom they had to teach how they wanted, they felt limited by how much they had to cover and they agreed that time spent complying with AchieveNJ was time that should have been spent on getting better.

Table 5.3: Impacts of the compliance burden of AchieveNJ

Requirement	**Primary Impact On**	**Nature of Impact**	**Examples**
Teacher observations, with pre and post	Principals, especially in district B because more untenured teachers	Cuts out time for less formal, more authentic interactions and “professional conversations”	I have 8 new teachers and 27 tenured teachers. That’s 78 observations and a total time of around 2 hrs each. That's roughly 160 hrs a year
Administering PARCC	All staff	Lost instructional time; teachers left feeling unappreciated	We sit in the hallways and give bathroom breaks if the teacher needs a bathroom break. Its really insulting. We sit in a chair outside the classroom the entire time. 8:45am to 10:30am. 2 days each week I sat in the chair.
Technical PARCC prep	All staff but mostly teachers in tested subjects	Time spent on how to take the test instead of content	We spend 30 minutes a day on technical prep (how to take the test);

			Spent way too much time this year just getting kids ready for the basic mechanics of the test. Estimate it at 2 periods a day for a week or 400 minutes or so. Meeting last monday to focus on tech stuff like how to log kids in etc. Was a regular faculty meeting. Entire meeting spent just on PARCC.
Documenting lesson plans	Teachers and principals, but mostly teachers	Distracts from time actually planning better lessons; reduces collaboration	Lesson plans take hours now. Used to be only an hour or two. Now, around 4 hours a week. No one pays attention to the time lost to changing transactional things, like revised formatting for lesson plans. Thats time not spent making things better. We purchased and spent money on books that have lesson plan formatting. I have to retype that. Can't just reference it. Rewriting every word of it identically. Typing exactly. But might pay for pdf converter. But all 5th grade teachers are doing the same thing. Manually typing it in to the lesson plan.
Documentation for observations and annual review	Teachers and principals, but mostly teachers	Frustration; distraction from more substantive preparation; reduces collaboration	The paperwork for announced observations adds more. Paperwork for annual is more inclusive. Have to give everything for your SGOs, with summations/justifications, PD plans (more expansive but not more actual learning). Whats the scale difference in the paperwork. In a five year

			period, overall scale is more than 10 times what weve done in the past. At least that much more time.
Documenting benchmark test results	Teachers	Distracts from more substantive things; reduces collaboration	Data entry clerks at times. Not only giving tests, but giving scantrons and and had to manually enter the results online
Document SGOs	Teachers	Distracts from substantive reflection	SGO is purely admin, does not drive/impact instruction. Forced, easily skewed goal setting
Covering all standards	District B teachers	Can't pace to match students' needs; may not be able to teach for mastery	Timing is an issue. Very little time to go back even though I know from the assessments the kids are behind. Does not give time to master the skills. Don't have time to do special projects. Have until March to do 44 standards. Last week was like Sophie's choice. If I waited til every kid got everything, I'd never move on.
Meeting about compliance requirements of NJ Achieve	All staff	Supplants other, potentially more substantive, PD	All our trainings in-house for the first year of TeachNJ were focused on meeting new regulations. Everything was learning how to implement. Training on new Info systems as well. And Training on PARCC. Some of it is transient but new staff will always need training. We missed a year of true PD for this.

5.3.2 Legitimacy and Authority

Legitimacy and authority are critical to teachers' success. Because their primary "clients" are present involuntarily, authority is important to fostering the cooperation necessary for teachers to maintain order and nurture students to academic and social improvement. (Bidwell 1965, Ingersoll 2003). At the same time, their authority depends upon their legitimacy, which is in turn tied both to the self-perception of their value and to the esteem accorded to the profession. (Lortie 1975, Grant 1989). As such, any reform that delegitimizes teachers or their profession risks eroding the authority that is critical to their success. Here I review evidence of the degree to which NJ's reforms are delegitimizing teachers in district A and B, finding enough evidence to suggest that legitimacy may be at risk. As with most other indicators, the evidence is stronger for district B teachers.

The evidence suggests three ways in which quantitative evaluation may be eroding teachers' legitimacy. The first two are tied to teachers' perception of their own value. The third comes from public perception. Each of the two effects on teachers' self-perception is closely related to conditions of failure identified in chapter 4. The first relates to the devaluation of educators' expertise discussed in detail in section 4.2. Teachers seem to internalize this devaluation and as a result perceive their personal role as less legitimate. The second is a consequence of being measured by an oversimplified metric. Each teacher's evaluation subjects the teacher to a new, more visible, highly simplified validation mechanism that does not always agree with their self-perception. Even where they intellectually understand that the number is not an entirely valid measure of their value, it nevertheless seems to affect their perceived value. The final

delegitimizing effect is external; the public discourse around evaluation and the very existence of test-based evaluation opened up a means of more directly questioning the value of individual teachers.

Teachers in both districts expressed frustration with the devaluation of their expertise. The evidence, however, suggests that frustration was not the only effect of devaluation. Several teachers' comments also suggest that they internalize the devaluation and feel less valuable. Here the comments were similar regardless of district. Sentiments such as "I feel disrespected [because I am] not really treated like an expert" from a district A teacher were echoed by district B teachers: "I don't recall anything [superseding] my judgment like testing does now." Others recognized that the great lengths policymakers went to in order to build an evaluation system would be unnecessary if teachers were trusted as experts: "Tremendous distrust in educators creates [the] need for validation." One area where district B teachers differed from their district A counterparts was in the sense of surveillance they experienced, amplifying the sense that externally-driven evaluation is an attack on their value. Two teachers reflected on this. One bemoaned, "teachers are under attack in everything we do" while another lamented how she felt "like... all the eyes are on you." In district A, one teacher lamented being "scrutinized" but the sense of a surveillance culture was not the same. That this might be different is not surprising given how different district A and B administrators communicated the role of evaluation. Despite the differences, the overall impact of feeling devalued and unappreciated was similar. One teacher in district A gave a concrete example that drives home just how much reform can change how important you feel, even in a district that goes out of its way to shield teachers. Describing her role during

PARCC testing, she dejectedly explained: "We sit in the hallways and give bathroom breaks if the teacher needs a bathroom break. It's really insulting. We sit in a chair outside the classroom the entire time. 8:45am to 10:30am. 2 days each week I sat in the chair." It's difficult to imagine other professionals being forced to sit and wait outside a room for a portion of their day, particularly to facilitate a means external evaluators came up with to assign a number to the professional's performance.

Similarly, teachers in both districts expressed not merely frustration with oversimplified measures of their performance, but also that the number, however limited in accuracy or consequence, had very real effects on their self-perception. Principals, at least in district A, described scenarios that corroborate this. District A principals lamented how much PARCC scores and their overall evaluation score affected their teachers' self-perception. One put it simply: "They think of [mSGP] as valuing the job they do even though it's only [worth] 10%."[33] Others described the stress teachers experienced from summative ratings that did not match their self-perception. "A lot of them were very upset" because "at the end of the year they got a number" that they would equate to being told "I am a B+ teacher" when "I am an A." "I want to cry with them and tell them it's just a number...but to them it is a number" they need to reflect their perceived value. While one principal acknowledged that she was seeing improvement in the classroom, she avoided attributed the gains to evaluation or the resulting stress. District A teachers' comments were consistent with their principals'. Said one, "I was slightly devastated [by my 3.49]. It's mostly just pride." Ultimately, in district A the issue

[33] In the initial launch of AchieveNJ, mSGP counted for 30% of a teacher's summative score. It was changed to 10% during the time period of my interviews. It returned to 30% in 2017-18 but was returned to 5% for teachers and 10% for principals in 2018-19.

is many teachers believing they are “highly effective” reacting negatively to being rated “effective” despite no real consequence of the difference.

In district B, there were similar comments about pride driving them to want high ratings: “If I wasn’t highly effective, I’d be absolutely distraught.” But the experience was more commonly of a different magnitude reflecting far stronger perceptions of failure tied to working with students with far higher needs. One 18-year veteran described why she was now counting down the years to retirement

> In the last two years I’ve never felt more pressure. To the point where you don't enjoy your job as much and feel like you're losing the battle. And no matter what you do, it's never good enough. And I think that, with this new process of evaluating us, when you thought you were doing well all these years and then one number is going to make you not so great anymore, it's discouraging... I'm wondering can it get worse. Can they take the joy of teaching out. I still love it. But you've undervalued. [You’re] not as valued as you were. You don't have the same respect. It's not just from administration. It’s from families too. No one’s valuing what you do anymore. Then they start playing the numbers game with you and you feel like, if I’m just a number... you wouldn’t want a number on a child either.

Another teacher echoed this sense of hopelessness caused by a number reflecting lesser performance than she perceived, describing how her colleagues struggle with feeling “that they are giving their 100% and they do see glimmers of hope in their classroom...but it doesn’t necessarily translate to the test.” An intellectual understanding of the limitations of the number does not seem to change the effect of the rating. “I know intellectually that it's not a good reflection of the kind of teacher I am. But I don't like having a smudge on my record. As much as I know its invalid and no matter what the number is it doesn't reflect what kind of teacher I am, it eats away at my soul. “ Thus, while the sentiment in both district A and B is similar, the degree is quite different: in district A, it makes teachers feel valuable but not highly valuable. In district B, it

fundamentally impacts their sense of self worth and creates a sense of hopelessness.

I found less evidence for reforms dramatically changing external perceptions of teachers' value. Two teachers, one in district A and one in district B, raised at least the possibility that external perceptions were changing. Said the teacher in district A, " Some parents really appreciate us but we get a lot of negative feedback from the community. Everyone has an experience with [teaching]. Everyone is a critic." The district B teacher was quoted above, lamenting that teachers "don't have the same respect...from families. No one's valuing what you do anymore." That I did not identify more evidence may be a product of my methods, which did not include interviews or other evidence gathering from outside stakeholders.

5.3.3 Motivation

The comments detailed above suggest another possible negative consequence of the factors leading to reduced control, flexibility, legitimacy and authority: reduced motivation. As described in section 2.2, the literature lays out how poorly designed accountability systems can impact motivation by devaluing the things teachers value, by devaluing their contributions to other outcomes like inspiring morals or a love of learning or by removing autonomy. Here I review the comments for evidence of the degree to which NJ's accountability measures are sapping motivation or otherwise harming morale. I find that motivation and morale are negatively impacted by both NJ's accountability measures and the compliance tasks that are necessitated by them. As with nearly all other impacts, I find the impact on motivation to be stronger in district B.

The evidence on motivation largely tracks that for legitimacy and authority. Three main factors are impacting motivation and morale: the feeling of being devalued as an

expert and the corresponding loss of voice that comes with it; compliance demands interfering with working with the children the way they believe is important; and the discouragement that comes from being measured by a metric that often disagrees with self-perception. As was the case with legitimacy and authority, the degree of impact is greater in district B than district A. District A staff are largely frustrated by lack of input and by being anything less than highly effective. District B staff, by contrast, are exasperated and discouraged with a palpable sense of hopelessness because the metrics and corresponding messaging from their administration are routinely bad despite their efforts and despite progress they see in the classroom. The result is that in district A, it is a nuisance that detracts some from motivation while in district B the impact is so great that it may be an existential threat.

District A principals noted that "teachers experience definite stress" from being measured and it causes "definite issues with morale." One example involved teachers from a school that was the lowest performing in the district on PARCC despite being high-performing overall. The principal said that regardless of their high performance, the teachers "respond to the fact that they are being outperformed" by other schools in the district. Said another principal, "last year I had tears. I had people break down." She noted that this was largely driven by a desire to be "perfect" in all facets of the profession and a feeling that complying with requirements for evaluation got in the way of all the other things they thought important, quoting one of her teachers, "how can I do everything perfectly, there's not enough hours to teach, to be there for the parents, to do all the communication, and do all this paperwork and get 4s in everything." The same principal, however, acknowledged that dissatisfaction did not rise to the level of "a

simmering revolt." Adding to the contrast with district B, this principal advised teachers not to overly stress about the results and to "take the weekend off and have a glass of wine." That said, at least one district A teacher expressed sufficient frustration with her lack of voice and compliance exercises interfering with working with children that she was considering leaving teaching altogether: "I feel guilty about it but I will leave teaching. I can see myself burning out. Being so frustrated. I already feel that way. My job satisfaction is very low. When it feels like its getting in the way of what you can do for children it gets really hard."

This sentiment of being at a breaking point was more common with District B teachers. It was common, almost ubiquitous, for teachers to use words like "demoralizing", "low morale" and "discouraging" and several used stronger, more internalizing language like "soul crushing" or "it eats away at my soul." This was partially for the same reasons as district A teachers' dissatisfaction, but the distinguishing factor in the degree seems to be how it is messaged and a sense that nothing they can do with change how they are perceived. Unlike in district A, district B teachers expressed that they are "always told they are not good enough" despite the fact that "teachers here are killing themselves." Another teacher put it similarly saying, "all you hear is that it's your fault. Everything is negative." These echo the quote in the prior section that teachers are discouraged because "they feel that they are giving 100%... but it doesn't necessarily translate to the test." As discussed more in the section on turnover below, this seems to be leading many teachers to at least consider leaving for easier teaching jobs or retiring early. This contributes to a sense that in district B the changes pose an existential threat more so than the mere nuisance they present in district A.

5.3.4 Collaboration

Collaboration is another commonly identified condition of successful instruction. The evidence here is mixed, but on balance, NJ's accountability measures do seem to be negatively impacting collaboration among teachers. Two district A principals alluded to substantive discussions about instruction being more difficult to talk about because they were now "sensitive" subjects. One teacher in district A expressed a similar sentiment, noting that the job has become "highly political." Another explicitly described the impact on collaboration, explaining that there is now "not much sharing across disciplines" both because of less time for non-compliance exercises and because "when people do get a chance to get together, [the discussion] becomes negative fast because everyone is under the same pressure." That said, at least one teacher argued that the CCSSs had a countervailing positive effect on collaboration: "Common core lets us share ideas across classrooms and limits reinventing the wheel," adding that they "can now do projects across the district." This furthers the argument that while summative evaluation often gets in the way, teachers welcome changes that directly inform instruction.

Given that district B teachers previously expressed a greater sense of politicization and feel more scrutinized by their administration, the impact on collaboration is unsurprisingly worse in district B than A. One teacher talked about how reforms have created a surveillance culture rather than a climate that is conducive to collaboration. "[It] changes the climate in here. You don't want to say anything to someone for fear that it gets back [to the administration], and you know it gets back. You have to be very careful. [For example] at union meetings, people won't speak up for fear that it gets back to administration." Another corroborated that reforms created a toxic

surveillance climate saying, “a lot of people fear administration. They don't want to be the one to take that first step. It's a risk you’re taking. There’s a target on you.” Another, longer-tenured teacher noted that “there is not as much group work” and speculated that she was “not sure if its because of what the state is doing to us” leading to an “every man for themselves kind of thing.” New Jersey’s accountability reforms did not create the surveillance culture in district B but they almost definitely exacerbated it.

5.3.5 Formative Data Use

While NJ’s reforms appear to be negatively impacting research-identified conditions of success, I identified another distinct impact as a participant observer in district C. The use of data for quantitative summative evaluation and the corresponding surveillance culture creates skepticism of data that makes it difficult to use data formatively. This is true even though staff on balance expressed a strong desire to use data formatively and recognition of its value for that purpose. Essentially, staff in district C were so fearful that the data would be used against them, they challenged even data that was explicitly formative rather than engaging with it to improve practice.

My role as the analytics lead in district C allowed me to witness firsthand how staff, especially principals, in district C engaged with data. I was an active participant in meetings to help principals choose measures for their evaluations and in meetings to roll out data tools that my team and I created with the express intent of making it easier for them to more easily use data to improve instruction. In the former, I witnessed a pervasive distrust of their central office evaluators. Principals simply did not trust that metrics would be used in a way that would be fair or would benefit them. This led some to prefer, for example, simple metrics like proficiency over more complex growth

metrics, even when presented with evidence of how much harder it would be for them to hit proficiency targets.

Interference with the ability to use data formatively was more problematic. One example is emblematic. My team and I created an early warning system (EWS) tool for principals, to make it easier for them to identify students in need of additional support. Some of the data in the EWS was related to their evaluations - for example, student attendance - but its use in the tool was entirely to identify students in need. At the training for principals on how to use the tool, we led off by assuring them the tool would not be used in their evaluation; we might track usage, but only to improve the tool, not as a means of oversight. Despite these qualifications, our training was almost entirely derailed by questions relevant only for summative purposes. Nearly every hand raised for questions was not about how to use the tool, but to express that they found something for which we should not hold them accountable. For example, many immediately identified students that had attendance issues they perceived to be outside their control and should not be included in their tool. When reminded that the purpose of the tool was not to highlight students for whom they would be held accountable, but rather those that need more help, principals expressed a lack of faith that any data would not be used against them. This issue arose every time my team and I presented formative data tools to staff to the point where our first sentence in every presentation was the qualifier that the data would not be used for accountability. It never stopped questions about how unfair it would be to hold them accountable for some data point contained in the tool.

After my experience in district C, I went back to my district A and B interview data for evidence of the same phenomenon. I found some evidence that this was the case,

though more conclusive evidence would likely require additional interviews, as this was not something for which I probed in the first instance. Several district A principals, for example, noted the extent to which their teachers spent time challenging their observation results. For example, "I've had teachers come with Danielson, tabbed out, with their lesson plan, saying here's why you need to give me a four. Some of the 15 minute post observation meetings are going to 45 minutes." This is directly related to its use for summative purposes as the incentive to challenge evidence is diminished in a formative environment. District B staff, as noted, routinely expressed their perception that everything might be used against them. "If you're not doing it, they getcha. [Summative evaluation makes it] easier for surveillance. It's all a gotcha moment."

5.3.6 Stressing students out

So far, all of the conditions for success have been about educators. Here I address the impact on students, as teachers routinely argued that testing was creating unproductive, sometimes unhealthy stress. The subject often arose as a follow-up to explanations of how much time they spent making students comfortable with the format of the test. For example, after one district A teacher without an mSGP described spending an inordinate amount of time on non-substantive test prep, I asked why she did it if it did not affect her evaluation. She responded, "We see the stress of the kids. It's unfair to them if we don't help them be comfortable and confident. Want them to feel like they can really show what they know." Others echoed the concern about students' stress. One described having to tell parents to "to stop talking about the high stakes to your kids 'cause they are coming in freaking out." Another lamented that the kids "had so much anxiety" and another that the "pressure trickles down to students." They cited at least one

negative consequence beyond the stress. “It creates a much more assessment driven-setting for students” and hurts the joy of learning. While this is an area that came up more in district A than B, district B teachers noted a similar stress: “kids don't know nothing’s riding on it. We tell ‘em but it still stresses ‘em out. They take it seriously even if we tell ‘em it doesn't affect them.” They also noted that in addition to impacting the joy of learning, the pressure makes it hard to teach for long-term understanding rather than short-term coverage.

5.4 State and district level implementation was a problem

The first three mechanisms of failure have dealt largely with flaws in the design of education reform and the theory behind it. However, it is entirely possible for a theoretically sound policy or intervention to fail even when designed well if it is implemented poorly. While implementation was not one of this dissertation’s original research questions, given how critical local knowledge and context were to implementation, I hypothesized after the fact that New Jersey’s education reforms would be poorly implemented. Despite the fact that my interview protocol was not designed to capture evidence of implementation issues, I found enough support for the hypothesis to merit a brief discussion.

One of the biggest implementation issues is the state’s choice of the metric by which to evaluate staff. As noted earlier, mSGP does not completely address issues of classroom composition and may be sensitive to students’ starting points. But there’s a bigger issue with the use of SGP and mSGP in summative evaluation. SGP and by extension mSGP are not valid for causal inference. As a result, SGP has possible unintended consequences in accountability systems. (Castellano, K. and Ho, A. p.100).

> "Student Growth Percentiles are often incorrectly assumed to describe an absolute amount of growth in a normative frame of reference. They are instead a relative metric in two ways, both with respect to the variables included as predictors and with respect to other students in the model. Group-level SGPs may be overinterpreted as value-added measures when they are not intended to support these inferences on their own." (p. 100).

NJ's DOE may have made the same incorrect assumption in choosing mSGP to evaluate teachers and principals.

A different issue with PARCC and mSGP is timing. To be most valuable, both test scores and mSGP would need be available soon enough to make timely decisions. Nearly all staff lamented that this was not the case. Instead, PARCC scores and mSGP often were not available until the following school year. Teachers do not receive test data in time or in a format that allows them to adjust instruction accordingly. Administrators do not receive it in time to make staffing decisions in the way likely envisioned by policymakers. A teacher lamenting how early the tests take place raised a related timing issue. "So I have less time to teach so you have more time to grade." The tests were given well before the end of the school year to provide time for results and results still came out too late to be useful.

Other issues were more administrative. Some principals complained of a lack of timely communication around PARCC. "The test is in two months and I'm just getting the manual from the state now. Seriously?" Other principals and teachers noted how the state's scoring systems did not allow for the use of NA's in FFT, even though they seemed routinely necessary, particularly for unannounced observations of classes that were not designed to cover all items in the framework.

Perhaps the biggest lesson is related less to the implementation of AchieveNJ specifically than to the nature of reform more generally. Teachers consistently lamented

the frequency with which changes were made and the corresponding lack of time they had to get good at anything. Said one teacher, “They like to change things without giving it a proper try.” Said another, “it will change in five years” even though it will take longer to see if it works. For longer tenured teachers, the effect of frequent reform was to blunt its impact. One teacher, explaining why she is not worked up over reform and felt comfortable mostly ignoring it, said, “being here long enough, I've seen a million things come and go.” A colleague echoed this, “I don't remember any time in my years of education where there wasn't something new being rolled out. Teachers have reform fatigue I think.” One teacher summed up the political nature of reform and its implications for teachers:

> “The constant change is what bothers me. I did one lesson and I want it to be better. [But] we can’t develop with it because everything keeps changing. [We] can’t make something be distinguished. If [the Governor] drops common core, that’s a lot of wasted work. How do I perfect it if I lose that. No one knows how much work it is to change it.”

This latter is particularly prescient given Governor Murphy’s efforts to undo some of Governor Christie’s initiatives several years after these interviews. Ultimately, this further validates Stephens (1967) observation that “every so often we adopt new approaches or new methodologies and place our reliance on new panaceas.”

5.5 Turnover

While turnover fits logically within sections 6.1.2 and 6.1.3 - it is both a form of resistance and undermines conditions of success - I treat it separately because it came up often and is such a strong leading indicator of failure. The other reason I treat it separately is, unlike other indicators of failure, turnover may be measurable relatively quickly. Using administrative data from district C, I was able to test a hypothesis that

arose from interviews in districts A and B. Like the other indicators here, interviews suggested that turnover would be a bigger issue for districts like district B than A. The literature also reflects this as a more general trend. As Ingersoll (2012a,2012b) and others have noted, turnover is far more prevalent in high-poverty urban districts than in high-income suburban districts. And as in many other fields, turnover, especially of high performing staff, is a costly problem. To the extent that reforms cause turnover, then, they are less likely to improve student achievement.

5.5.1 Evidence that reforms are making teachers consider leaving their schools or the profession

Here I focus on qualitative evidence from numerous staff in districts A and B. While teachers in both districts A and B indicated they might leave because of reform, the sentiment was much more common in district B. Similarly, while at least one principal in district A indicated that no one in her school had left as a result of reforms, there was no such counter-evidence in district B. Moreover, at least one district B teacher indicated that she was planning to leave at the end of the year in which I interviewed her and was planning to go to a district like district A. And in district B, the teachers that indicated planning to leave were higher performing according to their principals' and peers' reports. In nearly all cases, whether district A or district B, teachers indicated that reforms were directly reducing job satisfaction - for many of the reasons outlined in 6.1.3 - or were limiting their ability to serve students.

Two of four principals in district B cited high turnover with reform at the root. Both indicated the same compositional pattern: older teachers were retiring. One noted that her school had "lots of older staff turnover. I'm sure [it] was because of the changes." Another noted that she had seven teachers retire in the last two years in a

school with 30 total positions. Both principals indicated that this created problems with hiring and increased the burden of evaluations because new teachers get observed more often than tenured staff.

At least one high performing district B teacher I interviewed exemplified the trend of higher turnover amongst experienced staff. She was a 67-year-old elementary teacher who had intended to teach until she was 70. But now, she said, "I'm retiring in one more year. It's gotten too crazy. Can't I just teach? Leave me alone. Im done. It's not fun. It's too much."

That said, it would be a mistake to conclude that reform might only shave a couple years off the careers of older teachers. I spoke with several younger, high performing - according to their principals, peers and ratings (some showed me theirs) - teachers who were planning to leave. One, for example, said, "There are days where I felt I could do it forever. The minute nonsense [changes that]. We got emails today about the technical details of where the DRAs [for SGOs] go in the folders." One teacher in particular stood out as counter to the vision of reformers. She was young but experienced (15 years), highly rated, devoted to her students and extremely energetic. I interviewed her twice, six days apart. In the first interview, she spoke of the challenges and noted that her strongest colleagues might leave. But she was committed to staying. During the second interview less than a week later, she told me she was looking for jobs in the suburbs or at least in charter schools. Following was part of our first exchange, showing that she was frustrated but not yet ready to leave:

> There are some things that are very rewarding. Gratification hasn't changed. Within my four walls, what happens is magical. That part I love... All the stuff outside of my four walls...makes the load heavier and heavier. It's almost getting to the point where it's not enough. Early on, it's part of the territory. But now, the

> scales are tipping. You see a mass exodus of people. You know who will leave? The good people. Because they internalize this stuff. I internalize this stuff."

In our exchange 6 days later, the same teacher confided in me that she was actively seeking jobs in the suburbs or "at least" in a charter school that would value her efforts.

While it was less obvious in district A, there were still instances of dissatisfaction portending turnover. One principal, for example, noted that "this is the first year I've ever had staff have meltdowns. I've had staff ask how much of a financial hit it would be to retire early. It's mostly a quantity of work not the nature of the work. Joy of teaching is gone because of all these things." A few teachers in district A corroborated this. One difference between district A and B is that in district B, when teachers are not ready to retire, they suggested they might go to districts like district A. In district A, they talk of leaving the profession, as this teacher did: "I won't do this my whole life. So scrutinized and unappreciated. I feel guilty about it but I will leave teaching." Another district A teacher talked about her colleagues leaving at alarming rates because of reforms, highlighting that many were experienced or even teachers of the year. "Over a three year period, [we've] probably lost 8-10 teachers absolutely because of this. Two icons in this building alone, past teachers of the year, left this past year. " Thus, while district A may be better able to absorb turnover without large impacts on student outcomes, there is at least reason to think they are not immune from undesirable turnover.

5.5.2 Testing the hypothesis with quantitative evidence

In this section, I use administrative data from district C to test hypotheses about the relationship between turnover and reform raised by educators' statements in districts A and B. First, I ask whether turnover overall is increasing in district C. This does not directly indicate what the cause might be, however. For that, I use whether teachers had

an mSGP score in the prior year as a proxy for the "dosage" of reform to which a teacher is exposed and ask to what degree are teachers with mSGP more likely to leave than those without. Finally, recognizing that not all turnover is inherently undesirable - in fact dropping the lowest performing teachers is an express purpose of TeachNJ - I look at the composition of turnover, asking whether turnover is resulting in higher concentrations of highly rated teachers.

Overall there is little evidence in this data set to support strong claims on either side. Part of this could be a relatively small sample of teachers with mSGP and a lack of data on reasons teachers exited the district - there is no way to distinguish staff that left voluntarily from those that were terminated. Part of this could be that mSGP is a poor proxy for the "dosage" of reform, especially with it being such a small share of the overall evaluation score. The data that do exist, however, show little evidence of reform having direct impacts on turnover. Turnover did not increase in district C following TeachNJ or PARCC adoption and the relationship between having an mSGP score and leaving the district is weak, suggesting that being measured by test scores – the most controversial element of reform – did not meaningfully push teachers out. On the other hand, reform supporters have little to celebrate. Overall, the changes in the composition of the workforce were minimal. While the relationship between summative scores and exit was stronger than that for having an mSGP and exit, the dominance of teacher practice scores in the summative score along with the likely lack of fidelity with which teacher practice scores were generated suggests the composition of who exits is still determined more by principal discretion than "objective" measures.

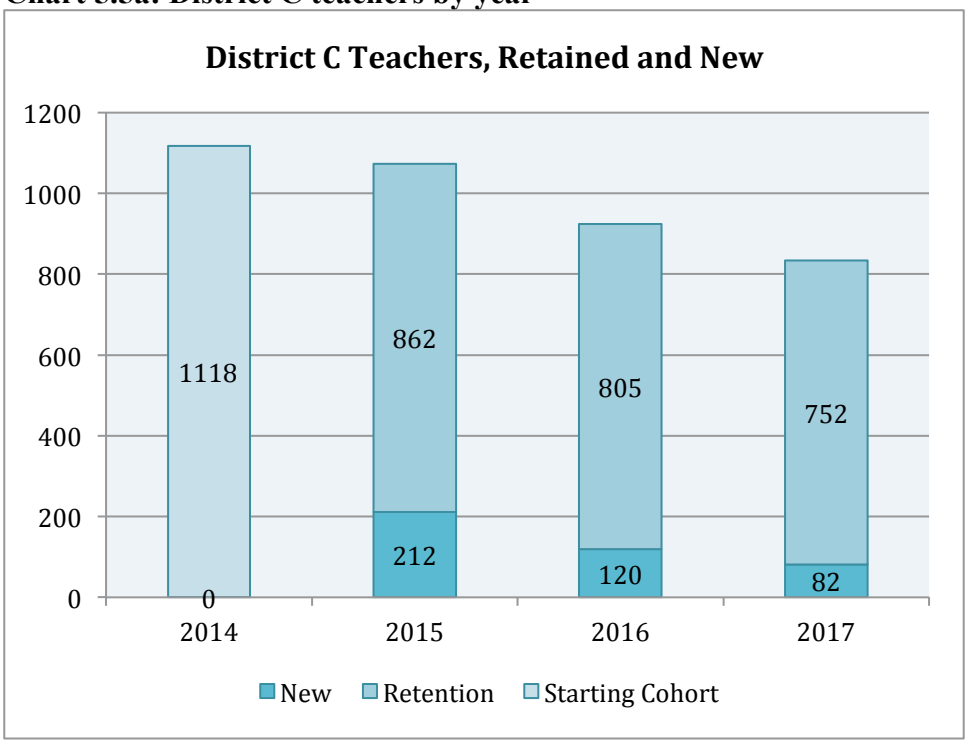
The Dataset

This section draws on an administrative dataset from District C. The data set covers four years from the 2013-14 school year (2014 in charts an tables) to the 2016-17 school year (2017 in charts and tables). It includes variables for school, grade, subject, teacher race/ethnicity and all available annual evaluation data: annual teacher practice score, SGO, mSGP score if available and overall summative score. As noted earlier, SGO's are teacher selected growth measures with relatively little weight in the summative score. They are an easily gamed metric seen as little more than an inconvenient compliance exercise. As such, I expect them to have little overall impact on turnover. mSGP scores, on the other hand, are associated with more of the issues educators raised than any other element of reform. Derived from students' scores on PARCC using an opaque statistical procedure, mSGP scores rely on a test many educators perceive to be an invalid measure of student performance and the most egregious attack on their expertise and discretion. If accountability reform is driving teachers out, mSGP might be the best single proxy to capture it. To these administrative variables, I added several generated variables including a binary variable for whether the teacher had an mSGP score and another binary for whether the teacher exited in a given year. I also added 1 year lagged variables for most variables to capture the conditions immediately preceding a teacher's exit or retention.

The total number of teachers passing through at any point over the four available years was 1,509. However, as shown in chart 5.5a the total number of teachers active in each year declined, matching a charter-growth driven decline in the district's student population. Despite the overall declines, however, there were non-negligible numbers of

new hires each year. Still, new hire numbers declined each year along with overall positions.

Chart 5.5a: District C teachers by year



Turnover is Not Increasing

Following Ingersoll (2012a), turnover in any given year includes any teacher that was in district in the prior year and is no longer in the district. While all teachers said they were planning on leaving, none said they were leaving imminently. Thus, we might expect turnover to increase over time as teachers retire or find jobs elsewhere. There is, however, little evidence that turnover is increasing over time. In fact, most of the evidence suggests otherwise. Overall turnover decreased both nominally and proportionally from 2016 to 2017, as shown in chart 5.5b. While turnover did exceed total position reductions, leading to additional hiring, it did so by less each year. As shown in chart 5.5c, the share of the workforce made up by retained teachers actually

increased from 2015 to 2017 from 80% to 90%. Conversely the share of teachers exiting in the prior year declined from 23% to 19% in the same timeframe.

Chart 5.5b: District C turnover by year

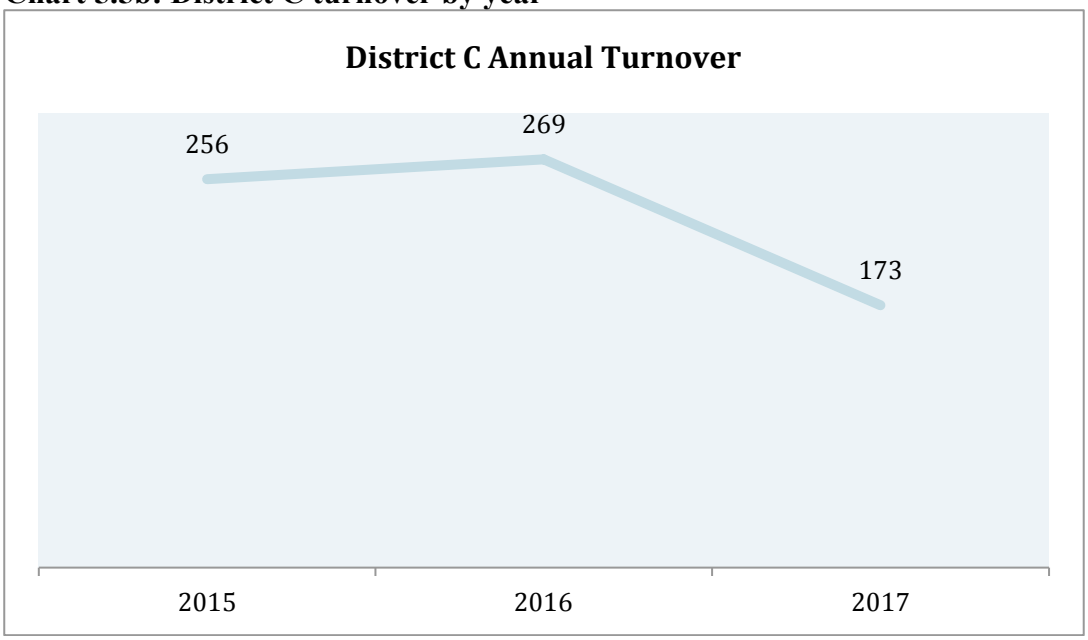
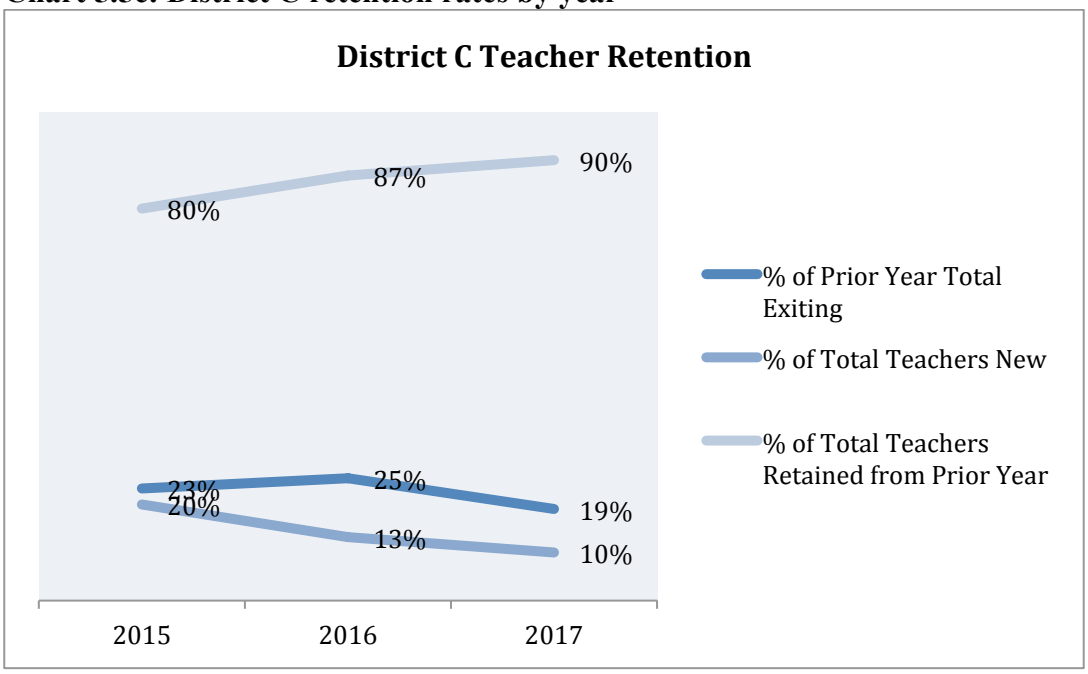


Chart 5.5c: District C retention rates by year



Does evidence suggest teachers are leaving because of reform?

The data set does not lend itself to a complete answer of this question. It contains neither reasons for exit – was it voluntary or involuntary? – nor a large sample of teachers with mSGP as part of their summative scores. Over all years, only 393 teachers had an mSGP score. Moreover, while the share of teachers with an mSGP score increased annually, it peaked in 2017 at 14%. So testing any hypotheses that require variation within the group of teachers with an mSGP was not possible. With these qualifiers, I sought to test the evidence that should be present if the hypotheses about reform driving teachers out were true. Relying on the fact that test-based accountability created the most pushback in districts A and B, I hypothesized that if reform is driving teachers out, teachers subject to the worst part of it - those for whom testing was part of their summative rating - should be more likely to leave.

There is little evidence that that is the case. Overall, as shown in chart 5.5e, only 15.7% of teachers left the district the year after having an mSGP, compared to 15.4% of teachers without an mSGP. Moreover, as more teachers received an mSGP score, a smaller share of those left. In 2016, only 1 teacher with an mSGP in 2015 left (see chart 5.5d). In fact, in both 2016 and 2017, the share of teachers with an mSGP score in the prior year that left the district was lower than the share of teachers that left without having an mSGP score, suggesting that, if anything, teachers with an mSGP score (see chart 5.5e) were less likely to leave, not more.

Chart 5.5d: District C turnover by year (teachers w/ prior year mSGP)

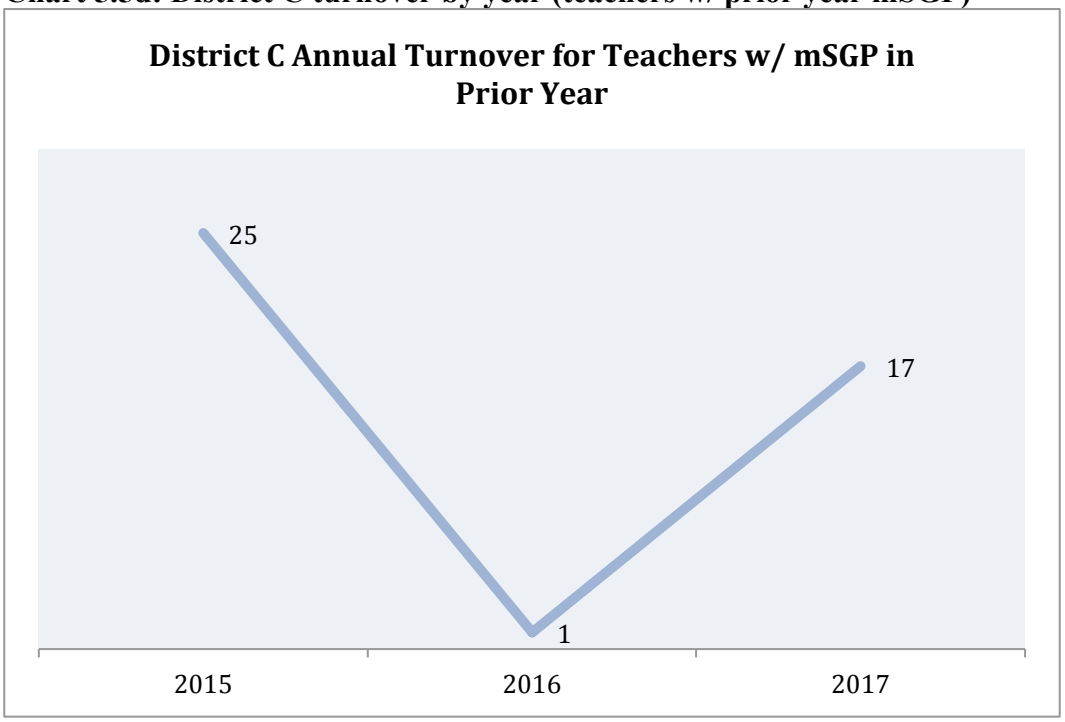
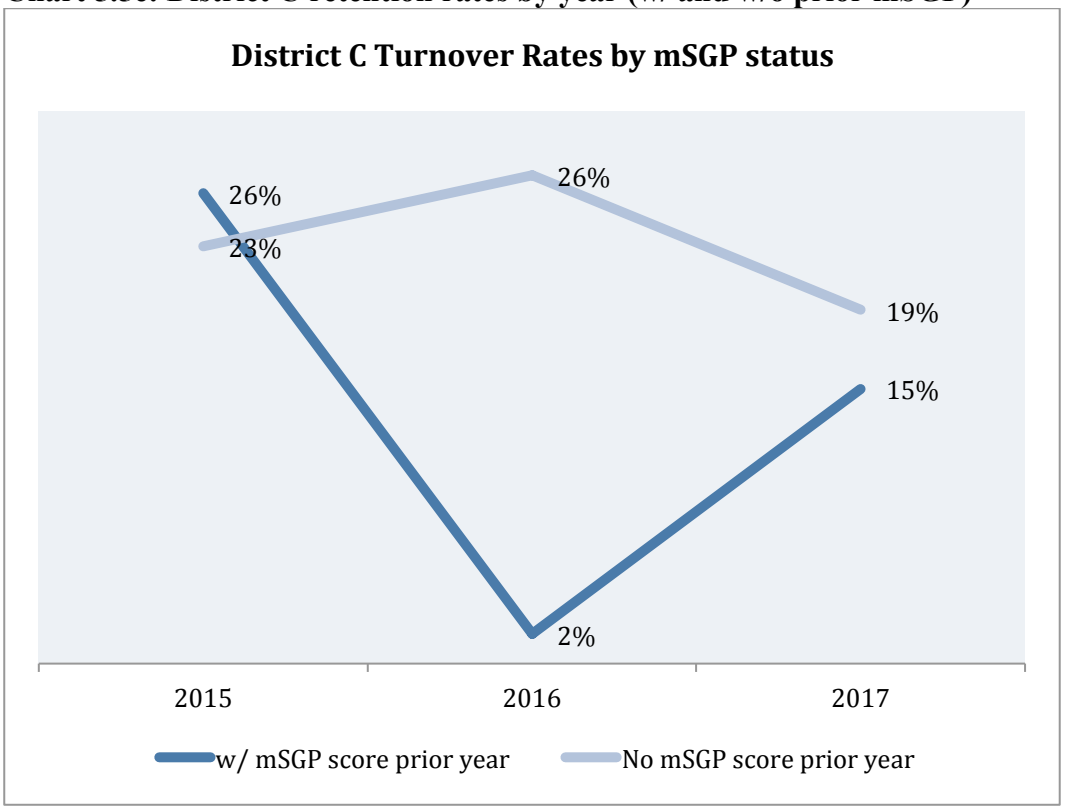


Chart 5.5e: District C retention rates by year (w/ and w/o prior mSGP)



The evidence that mSGP might be driving staff out is even weaker when factoring in that mSGP is also impacting summative scores and therefore, theoretically, involuntary exits. For example, mSGP scores were generally lower (.4 points on average) than teacher practice scores. And many teachers had substantial differences between the two. For example, a quarter of teachers with an mSGP score received an mSGP score over .75 points below their practice score. mSGP scores would therefore tend to drive down teachers' overall summative scores. As such, the existence of an mSGP score might increase involuntary exits and positively predict exit even if the existence of mSGP score had no impact at all on voluntary exists.

To more explicitly test whether the impact of mSGP on summative scores does in fact weaken the evidence for mSGP scores causing involuntary exits, I ran a logit with the binary for exit as the dependent variable and included a dummy for whether the teacher had a prior year mSGP as the primary independent variable. I also included a time variable (Time=0 at Year=2014) to capture any potential correlation between the year of analysis and both having an mSGP score and exiting.[34] An ideal model would have included all characteristics that are correlated with both whether a teacher has an mSGP score and with whether they are more likely to exit. Unfortunately, my dataset lacked at

[34] I originally planned to run the analysis using a fixed effects panel logit to address any time and effect invariant variables that are correlated with having an mSGP score and exiting. However, the intra-teacher variability in the three observed and included independent variables (had an mSGP score in prior year, prior year summative score and taught SPED prior year) was fairly limited given the large share of staff that never had an mSGP score and the year over year stability in summative scores and teaching SPED. At the same time, I would expect the time and effect invariant variables – namely gender and race - to be only weakly correlated with both having an mSGP score and exiting. Intrinsic motivation may be fixed and impact exit, but it is unlikely to be correlated with having an mSGP score. As such, the increased OVB from not using an FE model should be fairly small.

least one variable that meets these conditions while two other variables that meet these conditions were coded in the administrative data in such a way as to render reliable inclusion impossible. This suggests a fair likelihood of omitted variables bias. I address the degree and theoretical direction of this bias after detailing the model and omitted variables below.

To separate out the potential impact of the magnitude of mSGP scores – as opposed to their existence - on involuntary exit, I included the teacher's prior summative score as a covariate.[35] Likewise, I included whether a teacher taught a SPED class of any kind in the prior year, as this is also potentially correlated with both exit and having an mSGP score (turnover may be higher for SPED teachers and SPED teachers often teach smaller courses and therefore do not receive mSGP scores as often as other teachers).

Table 5.5a summarizes my analysis of omitted variables bias. The three variables that are included in my theoretical model and excluded from my logit model are class-size, grade-subject and year-over-year change in grade-subject. The first omitted variable is the number of students a teacher has in his or her classroom. This could theoretically predict their exit and is clearly correlated with having an mSGP score because there are minimum student counts needed to be eligible (teachers need 20 students in a given year or over consecutive years). I hypothesize a positive correlation between class-size and both exit and having an mSGP score. Therefore, I assume the omission of class size will cause my model to overestimate the impact of having an mSGP score on exit. The correlation between class size and having an mSGP score is likely fairly modest,

[35] Including the actual prior mSGP score as a covariate rather than prior summative would have restricted the sample to only teachers with a prior mSGP, eliminating the comparison group and making it impossible to test whether having an mSGP predicts exit.

however, because many teachers who teach untested grades and subjects also have large class sizes but no mSGP score. I therefore assume a moderate amount of upward bias.

The second omitted variable – subject-grade - is technically included in the dataset but coded in such a way as to render it too unreliable to include. [36]. Theoretically, the subject a teacher taught and/or the grade level might be correlated with both having an mSGP and exiting. It is not clear, however, what the direction of the correlations might be. For example, are teachers in subjects like 4th grade elementary more or less likely to leave than their counterparts in high school art *independent of* the effects of the former being a tested grade and subject? Or less? Conversely, might the difference be reversed for teachers of untested students in PK-2? This ambiguity suggests that while the direction of any bias is unknowable, the correlations might be weak and therefore the degree of bias small.

The third omitted variable – year-over-year change in subject-grade – would have to be derived from the second and was therefore omitted for the same reasons. Whether a teacher changed grades or subjects would theoretically be positively correlated with both exit and having an mSGP score. Teachers might be more likely to leave if they have to teach to a new grade or subject and it is possible that the district was more likely to switch teachers into tested grades and subjects given increased focus on those grades and subjects. As such, I assume the omission of this variable, like the omission of class-size, would cause my logit model to overestimate the impact of having an mSGP score on

[36] My dataset did contain variables for both grade level and position. However, the coding was extremely granular and inconsistent both within and across years. For example, the majority of teachers were coded with multiple grade levels and year-over-year coding of subjects was inconsistent. While I attempted to manually recode these, I simply was not confident my judgments were reliable.

exiting. That said, the number of teachers changing subjects-grades in this way was likely fairly small. I therefore assume the overall degree of this bias will be small.

Table 5.5a: Omitted variables bias in panel logit model

Omitted Variable	**Direction of Bias** Model wil _____ the impact of having an mSGP score	**Degree of Bias**
Class Size	Overestimate	Moderate
Grade-Subject	Unknown	Small
Change in Grade-Subject	Overestimate	Small
All three combined	Overestimate	Small to moderate

Thus all three variables might introduce some bias in the model but I assume that bias is small to moderate. At the same time, that modest bias is likely to be upwards. That is, I assume my model will overestimate the impact of having an mSGP score on turnover. As the results in table 5.5b show - and contrary to my hypothesis developed from the qualitative interviews - having an mSGP is associated with a marginally significant *decrease* in the likelihood of leaving the following year. Calculating the marginal effects at the mean summative score and modal SPED binary and Year 3 (2016), a general education teacher with a summative score of roughly 3 is about four percentage points less likely to leave if they have an mSGP than if they do not. Given the small number of teachers each year with mSGP and the aforementioned likelihood of omitted variables, this is unlikely an accurate estimate of the true impact of having an mSGP on turnover. I cannot conclude that teachers are in fact 4 percentage points less likely to leave if they have mSGP. However, because it is likely that any bias would cause us to overestimate the effect, this provides fairly strong evidence against the hypothesis that the existence of mSGP is driving up turnover.

Table 5.5b: Predictors of exit using logit

VARIABLES	Model 1 Coefficients Exited	Model 2 Coefficients Exited
Teacher had mSGP score	0.0338	-0.363**
	(0.171)	(.180)
Teacher's prior year summative score		-1.511***
		(.140)
Teacher taught SPED prior year		-0.128
		(.116)
Time (years since 2014)	-.212***	-.168***
	(-.049)	(.062)
Constant	-1.290***	3.378***
	(-.102)	(0.404)
Teacher-Year Observations	4,527	2,860
Unique Teachers	1,509	1,367

Robust Standard Errors in Parentheses
*** p<0.01, ** p<0.05, * p<0.1

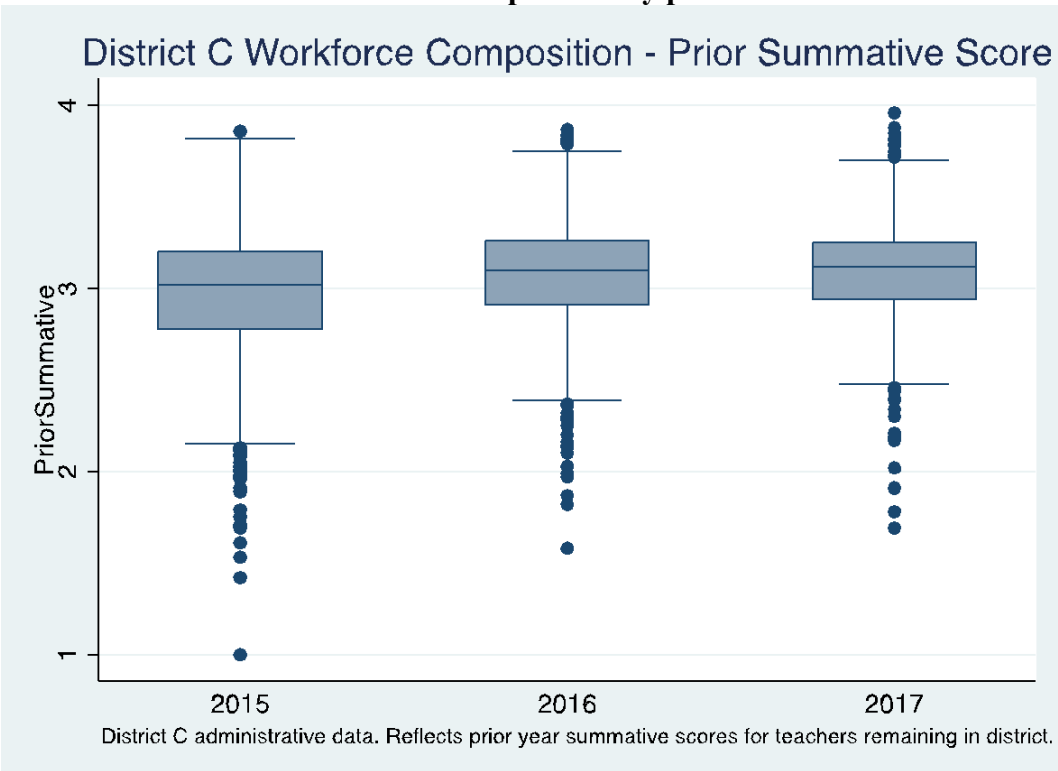
Is reform impacting the composition of the workforce?

So far the evidence from district C is inconsistent with dire hypotheses about reform causing a mass exodus of teachers. The limited evidence shown above suggests that turnover is not dramatically increasing beyond that driven by declining enrollment. Similarly, the data does not suggest mSGP is meaningfully increasing involuntary exits. The final question, though, is whether reform - and summative evaluation in particular - is having any impact on the composition of the workforce. Reform proponents argue that the changes should lead to increasing concentrations of higher performing teachers. Opponents, and several of the teachers in districts A and B, suggest the opposite is likely. The evidence is not strong for either argument.

As charts 5.5f and 5.5g show, when measured by either the prior years' summative scores or the more objective prior years' mSGP, the composition of the teaching workforce did not change dramatically either way through 2017. Summative

scores show a slight upward shift with some compression around the median suggesting that the district may be successfully removing the lowest performing teachers. However, given the limited fidelity with which principals were likely scoring observations and the fact that teacher practice made up the majority of the summative score, I hesitate to draw strong conclusions from this.

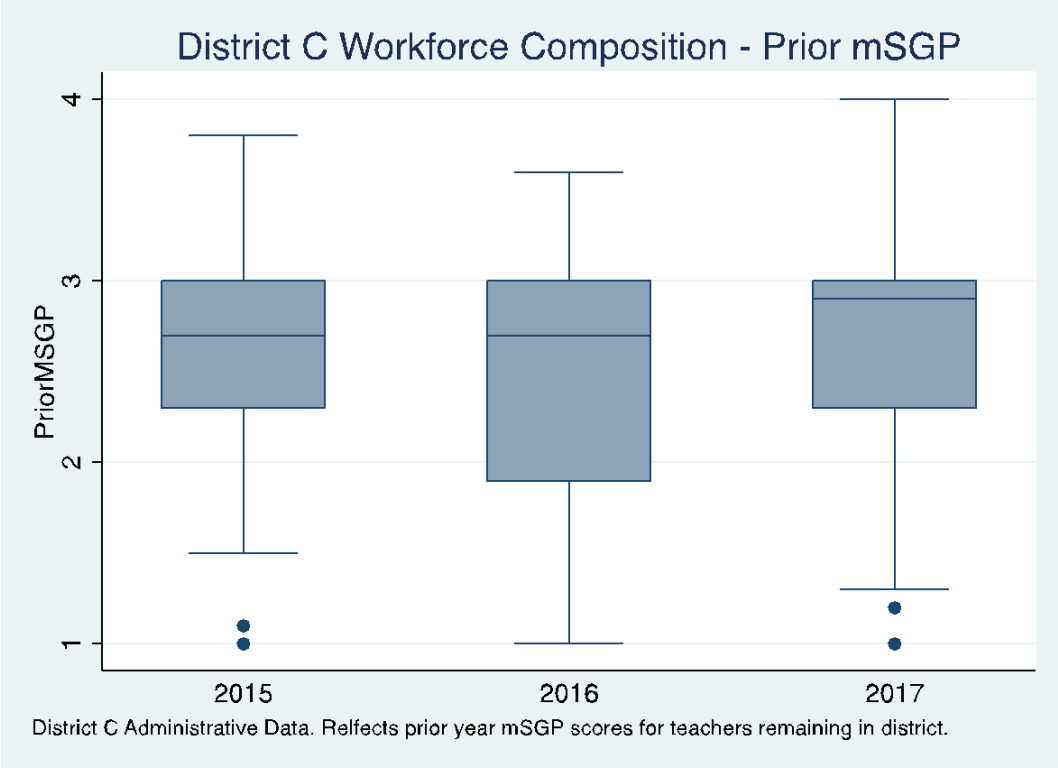
Chart 5.5f: District C workforce composition by prior summative score



While mSGP scores have their own limitations, not least that barely more than 1 in 7 teachers had them and they are inherently relative, they are not subject to the same subjectivity. They may therefore be a somewhat more reliable indicator, at least from the perspective of reform proponents. Chart 5.5g shows the distribution of current teachers' prior mSGP scores for each relevant year. The median climbed slightly between 2015 and 2017, from 2.7 to 2.9, but the 25th and 75th percentile scores stayed flat at 2.3 and 3 respectively. The 10th percentile score dropped slightly from 1.7 to 1.6 while the 90th

percentile score dropped slightly from 3.6 to 3.5. All of this suggests that reform is not meaningfully driving out either the lowest performing teachers or the highest performing teachers.

Chart 5.5g: District C workforce composition by prior mSGP score



The evidence regarding turnover is therefore weak for the arguments of both proponents and opponents of reform. At least in district C, it does not appear to be meaningfully reshaping the workforce. In that regard, the evidence may be fairly strong for another hypothesis raised by the literature and the longer tenured staff in districts A and B: like prior reform efforts, current efforts may be a really costly way of changing very little.

6 Conclusion

6.1 Explaining education reforms' consistent failure to change the trajectory of educational progress

This dissertation proposes an answer to a fundamental question about public education: if public education is constantly being reformed, why are the results always so consistent? Given the breadth of this question, I drew on broader policy literature to develop a framework, then tested that framework against evidence from case studies of three school districts in New Jersey, an exemplar of the national trend in education reform. My findings suggest this framework offers a promising answer to the mystery.

The framework in this dissertation derives from an argument that education reform is most fundamentally a case of a state attempting to govern human behavior to achieve a policy goal. This situates education reform within Foucault's theories of governmentality and suggests the best way to understand education reform is to consider the technologies government is using to achieve its ends. A common feature of all such technologies is a need for simplification. As not all technologies fail to meet the goals of their designers, simplification alone is insufficient to explain education reform's disappointing results. Building on Foucault's theories, case studies from Scott (1998) and Mitchell (2002), among others, offer a deeper explanation. Legibility efforts, to use Scott's term for these types of reform, fail when they oversimplify a highly complex social or natural process, centralize expertise at the expense of situated knowledge and ignore context.

Evidence from the districts studied in this dissertation shows that New Jersey's education reforms clearly satisfy these three conditions. I found ample evidence that

reforms both as designed and as experienced oversimplified the highly complex social process of educating students, centralized expertise in state bureaucrats and technical experts at the expense of situated knowledge and ignored contextual factors that are critically related to educational outcomes.

The most fundamental simplification comes from focusing almost exclusively on individual performance of teachers and principals. But the means of simplification are also highly problematic. Teachers and principals volunteered dozens of examples of what test scores and practice scores, both derived from instruments that capture only a share of the construct of interest, do not measure.

The evidence for delocalization of expertise was equally clear. Most notably, the teachers' local expertise - intimate knowledge of students earned from seeing them 180 days a year for several hours a day – is devalued in favor of the expertise of the creators of PARCC and the statisticians who use the results to generate student scores and teacher and school aggregates. This was an area particularly salient to teachers, who experience the devaluation of their expertise as a devaluation of their role.

Finally, there is clear evidence that New Jersey's education reforms ignore context. No adjustments are made – beyond the limited controls implicit in student growth percentiles – to accommodate the differences between districts A, B and C. These differences highlight the degree to which education reform ignores critical out of school factors that are in many cases far more predictive of student outcomes than anything individual educators do. Teachers in particular internalized this, knowing from experience the factors – like family instability - that dominate their students' experience but are ignored by reform. Another key finding from the assessment of context is that

district A teachers are well supported by their administration and insulated from state interference whereas district B teachers experienced the opposite; their administrators passed on and sometimes amplified the outside pressure. Overall, contextual differences showed up repeatedly to highlight how different the experience of reform was in district A and B.

With education reform clearly satisfying the conditions that spelled failure of other legibility efforts, the next question was by what mechanisms do these conditions lead to failure? Relying on the general policy literature and the education literature, I derived four mechanisms that connect the conditions of failure to the results. Policies that satisfy the three conditions lead to failure because they will be based on mischaracterizations of the problem, they will engender resistance, they will undermine the conditions of success and they will be poorly implemented. All four were well represented in the districts studied here.

New Jersey's education reforms mischaracterize the problem in that they focus on the incentives of individual actors. This is implicit in accountability-based solutions, as individual accountability is designed to align the incentives of individuals with those of policymakers. There is no evidence to suggest that educational outcomes are lacking because educators lack motivation to do better. Instead, the evidence suggests that educators are highly motivated by a desire to serve their students. The more likely problem, as evidence by educators' strong preferences for formative tools – including CCSSs – is that educators on balance do not know how to improve outcomes. Given this, no amount of accountability is likely to change the results dramatically.

The evidence of resistance was similarly strong, though took on different character in districts A and B. In district A, resistance manifested itself as loose coupling, with administrators insulating teachers from being overly burdened by state pressure. This was distinctly not present in district B. Instead, district B teachers resisted their own administrations' implementation efforts in subtle ways. For example, they opted not to follow curricular or pacing guidelines.

Reform similarly undermined the conditions needed for successful instruction. I derived four conditions in particular from the literatre: control and flexibility, legitimacy, motivation and collaboration. I also found evidence for two more, one of which – the use of data to inform and improve instruction - was previewed by Rowan and Raudenbush's (2016) finding that there is an inherent tension between formative and summative measures. The other was student mental health. I found evidence that reform undermined all six conditions. The administrative- time - burdens of the accountability regime limited teachers flexibility, the devaluation of expertise had a delegitimizing effect on teachers, motivation in district B was suffering under the "soul crushing" negative tone, and, again in district B, collaboration was limited because of the surveillance culture. In district C, I experienced principals actively resisting using data tools formatively because of the mistrust created by the surveillance culture. And teachers in both districts A and B spoke sadly about students experiencing high degrees of stress from standardized testing.

The evidence for poor implementation was somewhat weaker owing largely to the fact that I did not make an effort to capture implementation data in detail. However, teachers complained of insufficient notice of testing processes and other issues echoing

the findings of an FIU study of TeachNJ's implementation (South Florida Education Research Conference 2014).

Perhaps most glaring in all the findings was the degree to which the same policy was experienced differently in districts A and B. In district A, reforms were an inconvenience, frustrating but not demoralizing. In district B, the sense was that reforms posed an existential threat. This likely reflects both the different relationship between administration and teachers in both districts along with the different scale of the challenge the teachers faced - as district A teachers acknowledged, their students would likely be fine regardless of what the teachers did.

The weight of the evidence thus supports the utility of this framework as an explanation for why public education has had such consistent results in the face of constant reform. Policies that oversimplify complex issues, delocalize expertise and ignore context have limited prospects. Education reform, focused currently and historically on individual accountability, satisfies all three conditions. It is therefore unsurprising that it has failed to change the trajectory of educational progress.

6.2 Potential broader implications of the current round of reforms

Given the resources allocated to education reform, the lack of dramatic success of education reforms might by itself be enough to caution policymakers against making the same mistakes. However, my findings also suggest that legibility efforts may have ramifications well beyond their lack of success. Here I address three broader potential consequences of simplification efforts that may be implicated by recent education reforms. First, efforts to bring complex processes under the control of a central authority often result in the consolidation of power. This goes hand in hand with the centralization

of expertise at the expense of situated knowledge. Second, and relatedly, simplification efforts often create or perpetuate injustice. Finally, efforts that simplify through calculability tend to prioritize that which is counted at the expense of that which is not. Because broader consequences of reform are more attenuated and therefore more difficult to find evidence for, this section, is largely about generating hypotheses for future investigation and raising normative questions for public discourse.

Table 6.2 Broader social consequences of the conditions of failure

Condition	Broader Social Consequences
Oversimplification	● Prioritization of that which is counted over that which is not
Devaluation of Local, Situated Knowledge	● Consolidation of power ● Creation and perpetuation of injustice
Ignorance of Context	● Creation and Perpetuation of Injustice

6.2.1 Consolidation of power

To this point I have defined education reform as a failure because it has not resulted in the dramatic outcome gains proponents have publicly sought. Here I acknowledge the power dynamics that suggest that the publicly stated goals are not necessarily the goals of all stakeholders. As such, I acknowledge the possibility powerful stakeholders may not consider education reform a failure. Like all policies, education reform has winners and losers. A key question when considering potential broader implications of education reform, therefore, is who might be winning despite – or, more cynically, because of – the lack of dramatic changes in educational progress, particularly in urban areas. Because this is beyond the scope of my evidence gathering evidence, I raise the question here without directly answering it. Instead, I consider the means by

which legibility efforts in general consolidate power and look for preliminary evidence in the experiences of the district staff I interviewed.

Legibility efforts may consolidate power in at least three ways. First, they centralize and shift expertise and, as a corollary, consolidate power in those with particular expertise. In Mitchell's (2002) mapmaking case, for example, the new, more "rigorous" cartographic processes elevated the expertise of British mapmakers with particular skills. At the same time, the local surveyors lost a great deal of power. Their expertise became a mere footnote in the map. A similar phenomenon took place in Scott's (1998) mono-cropped forests, as the centralization of expertise allowed far less skilled laborers to work in the forest, reducing the power of the previously necessary ground-level logging experts.

The second way simplification efforts consolidate power goes beyond technical expertise. Efforts to make the world legible to a central authority are by their nature a means to the end of giving a central authority control over something over which it previously lacked control. The technologies of governance with which Foucault was concerned exist to make that which was previously too complex, too remote or too large to manage manageable. Mitchell's (2002) mapmaking case and Scott's (1998) monocropped forest case are again instructive. In the former, the map gave a central bureaucrat the ability to monitor the entire area of Egypt from his office, granting his office unprecedented power to control the legal, financial and social relationships represented by the markers on the map. Likewise, scientific forestry gave a central bureaucrat power over the entire forest from his office.

Finally, simplification tends to be hegemonic, enhancing the power of an already dominant class that has the ability to define the problem and, by extension, the objects to which policy should be addressed. Apartheid South Africa offers a particularly jarring example. Governmental technologies gave those already in power far greater capacity to control the population. They could, for example, more easily ensure that members of certain groups were prevented from receiving the "privileges" reserved for other groups. Mitchell's (2002) analysis of the tropic definition of the food shortage problem in Egypt is a subtler example. The tropic definition served the purposes of both powerful international agencies, by ensuring that the object of analysis was seen as separate from the experts that were analyzing it, and of Egyptian elites, by directing attention away from issues of equity that manifested in what looked like a food shortage.

The evidence from the case studies provides at least some hint at possible power consolidating effects of the latest round of education reforms in New Jersey. Most obviously, reforms have likely enhanced the power of those who design and score tests, analyze the results and sell solutions that most improve test results at the expense of teachers and principals. Similarly, reforms appear to have granted state bureaucrats far more control over what happens in district B's classrooms than they might have had before. This shows up, for example, in the degree to which teachers experience and respond to the pressure to focus on test scores. While district A teachers experienced some pressure here, district B teacher's experience more, possibly because the dependence of the district on state funding and the lower test scores of its students makes district B less likely to insulate its staff from outside pressure.

More speculatively, there is evidence of hegemonic power dynamics in what is ignored by education reform – the outcomes that have been simplified out and relevant context that has been bracketed. Much as the narrow definition of the food shortage problem in Egypt served select stakeholders' interests by avoiding complications of social policy and class, it is plausible that focusing on individual educators serves similar interests by avoiding similar complications. Among these complications are social issues of race and class, not to mention structural issues with the economy and thorny administrative challenges like reforming the institutions that govern education to align better to modern goals.

Because we can quantify the explained variance in educational outcomes, we can actually quantify how much has been ignored by education reform. This allows for neatly framed questions, the answers to which undoubtedly implicate power dynamics. What happens in schools explains only a small share – estimates range from 15-30% - of the overall variance in educational outcomes. If the large majority of the variance in educational outcomes is explained by factors outside the school, why is education reform so narrowly focused on actors inside the building? Beyond that, why is public education deemed both source of the nation's greatest threats and solution to its greatest problems? These in turn raise questions, unanswered here, about whose interests this apparent inconsistency serves. I make no effort to speculate on individual's motives, and leave space for the possibility that the results here are the consequence not of intentionality but of the fallibility of those wielding power with otherwise benign intentions. At the very least, however, we should wonder why anyone would expect annual gains of more than 3-4% by targeting at most 15-30% of the issue.

6.2.2 Creation and perpetuation of injustice

A related idea to the tendency of legibility efforts to consolidate power is their tendency to create and/or perpetuate injustice. Those who lose power as a result of its consolidation frequently lose more than just power. In some cases, such as the use of governmental technologies to effectuate apartheid policies, this perpetuation of injustice is patently obvious. The injustice is also obvious in Scott's (1998) case study of scientific forestry. As Scott notes, "(t)he monocropped forest was a disaster for peasants who were now deprived of all the grazing, food, raw materials, and medicines that the earlier forest ecology had afforded." Likewise, Mitchell (2002) argues that local farmers and landowners lost more than power as a result of the great land map. "The map helped to constitute and consolidate the new institution of private property and the forms of debt, title, dispossession, and violence on which it depended." (p. 83). The maps allowed the Debt Commission to transfer hundreds of thousands of acres of land from local inhabitants to private investors. John Dewey suggested that test-based education reform may perpetuate or create injustice nearly 100 years ago. Dewey argued that the sorting of students via tests amounts to a scientifically legitimized caste system. (1922a).

The evidence that NJ's education reforms are creating or at least perpetuating injustice is really a reflection of a pattern identified in nearly every negative outcome analyzed so far. The fact that, in nearly all cases, district B experiences the negative outcomes more acutely than does district A suggests that, at the least, reforms are perpetuating the injustices that have led to such different starting points for students in each district, and at worst, may be amplifying those injustices. While the laws and regulations are, on their face, the same, their application in different contexts clearly has

different impacts. In a place that is dependent on the state for funding and enters with low test scores, the pressure to change and the discouragement from routinely seeing poor results on the centrally chosen measure of success is far higher than in a district that gets most funding from its families and whose students perform well on tests almost regardless of what happens in the school building. The application of a system that has central administrators evaluating principals and principals evaluating teachers is far less risky in a place where all parties trust each other and feel supported than in a place where distrust is common and educators feel like administrators are out to get them. The potential injustice is that reforms are more likely to deter progress in district B where improvements are most needed.

6.2.3 Prioritization of that which is counted at the expense of that which is not

Finally, schemes based on categorization tend to prioritize that which is counted at the expense of that which is not. (Bowker and Starr 1999). In Scott (1998), the implicit categorization of trees as commercially valuable led to the elimination of all plant life in the forest that was not so categorized. In education there is clear potential for test-based accountability mechanisms to cause the deprioritization of other goals. My findings highlight just how much is excluded by reform measures. The risk is that, by attaching these measures to accountability, other socially valuable outcomes will get deprioritized. "When test results become the arbiter of future choices, a subtle shift occurs in which fallible and partial indicators of academic achievement are transformed into major goals of schooling…" (Glaser 1987, p.166).

The starting point for analyzing how much of risk current reforms pose is the outputs they prioritize.. The legislation and regulations are designed to "raise student

achievement" by "improving instruction." These are the two objects the accountability system aims to measure. Both, however, are complex constructs and measuring them requires highly political choices; they are impossible to measure without narrowing down to simplified proxies. For student achievement, the proxies are math and ELA test scores - more specifically, PARCC scores. Other subjects continue to be tested, but only ELA and Math are used in teacher and principal accountability. Instructional quality is likewise proxied, in most cases - including all three districts covered in this dissertation - by FFT. The question asked in this section can therefore be broken down to two more specific questions. To what degree does the accountability-backed measurement of math and ELA scores lead to devaluation of other subjects and/or goals of schooling? To what degree does the accountability-backed measurement of the items contained in FFT lead to the devaluation of other forms of instruction. My hypothesis was that test scores lead to more devaluation than FFT despite FFT having more weight in the accountability system because of the greater degree of oversimplification involved in using a single score from a single instrument farther removed from the judgment of local actors. The evidence supports this hypothesis.

Nearly every principal and teacher referred in some way to having to choose math and ELA over other priorities. As with nearly all consequences addressed in this dissertation, the degree of de-prioritization differed between district A and B. For example, one principal in district A lamented not being able to "do anything that can't be standardized" but her colleague spoke fondly of retaining a great deal of discretion over service learning and being able to provide healthy food options to students. That said, the same principal highlighted the tension, expressing concern that those things and the

things they contribute to are not valued by a system that measures only test results. "I make the argument that we are stronger than what NJASK shows because of our sense of community, level of parental involvement, satisfaction, our ability to differentiate instruction. We really educate the whole child here." Another principal made a more explicit connection, noting that testing time meant kids in her school got limited health and family life classes.

This is also an area where the sensitive subject of district efforts to align curriculum across buildings in district A came up. The need to align was driven by public pressure spurred by publication of school-level test results. As one principal reflected, "a big problem the last few years is that members of the community have looked at our [test] scores and started comparing schools." One outcome was the loss of a well-received program in a school. As told by that school's principal:

> We used to have a very successful multi-age program here [with] first and second graders in the same classroom. To make a long story short, we decided to get rid of it. So here it was ingrained in the fabric of what we do. It was a reason why we're successful. Huge buy in from parents and teachers. The community that's built when children spend two years together with the same cohort, same teacher, closely knit community among parents that's developed, the benefits were incredible.

Ultimately, the inability to measure the benefits of this different vision in a way that could compete with a public narrative of schools underperforming their peers on tests led to the loss of this vision. In a district like district A, the loss to the students may not be devastating, especially in the short run. But it is reasonable to wonder if there are ecosystem benefits being lost that may later prove to have been critical to the resilience and/or success of public education more broadly.

Teachers in district A generally agreed with principals that there were some other aspects of education deprioritized in favor of ELA and math, but nothing rose to the level of an existential threat. One teacher explicitly noted that math and ELA were prioritized at the expense of other subjects. Another noted that testing in ELA and math drastically reduced chorus time. And another noted that time and money spent on PARCC testing reduced professional development and pushed other subjects, including social studies and the arts "to the backburner." Their frustration, it should be noted, was more about the role of test-based accountability in forcing a narrow focus than the specific subjects. My sense was that teachers would have been equally dissatisfied if social studies and art were tested and prioritized over math and ELA.

Unlike their district A counterparts, district B principals expressed no concerns about other subjects being deprioritized. Said one principal, "the idea of it pushing out art, music, etcetera is ridiculous." It is difficult to infer whether this was because district B's administration chose principals who were sympathetic to the accountability environment in a way that district A's administration did not or whether there really was very little devaluation of other subjects. However, the fact that district B teachers, as described below, were more concerned with this than their district A counterparts suggests it might be the former.

District B teachers were highly concerned about the devaluation of other educational outcomes and to some extent this seemed to relate to the higher degree of pressure they perceived from their administration and the state. Two of the more experienced teachers, interviewing together, suggested that the threat of cutting funds was a real threat to services for children and families. They argued that their professional

integrity is what protected their classes from pressure to abandon broader goals of schooling and were concerned that newer teachers would not have the same willingness or desire to resist outside pressure. “We won’t be concerned with educating the whole child and the types of citizens we raise to problem solve...and [we won’t be concerned with] social-emotional behaviors.” Their colleague made the connection to testing and accountability more explicit, saying, “they have in a way limited my ability to do my broader duty. I want to know if students are making the right choices when we are not watching them. [I] want to help them be better respondents to their environments.”

Other teachers lamented the lack of time for other subjects or fun projects. Said one, “[we] rarely even have time to hit those subjects like science, social studies, health. Kind of gets pushed to the side because there is so much emphasis on LAL and Math.” Another said she struggles to balance teaching test-taking strategies needed for PARCC with real world lessons like “how to count money at the store.” Their colleague abandoned the “fun stuff,” noting, “When I first started, I had arts and crafts. No more.” One teacher in particular explained the trade-off decision directly, explaining, “We have to pick and choose. Will the skills translate to stuff we don't need for test? If it does not translate to the test, it’s low on the totem pole.” Finally, echoing the issue raised by district A principals, another district B teacher noted that her school, which was higher achieving than the others in district B, had to abandon their curriculum in the name of alignment. As with district A, district B educators did not blame either common core or observations for the need to prioritize math and ELA. The challenge came from test-based accountability.

6.3 Recommendations

This framework thus has a promising ability to explain the consistent results of public education in the face of constant reform efforts. It also has potential to raise important questions about the broader impacts of reform that are otherwise poorly adjudicated in public discourse. But the utility of the explanatory framework rests in the degree to which it helps future policymakers avoid the same mistakes. I conclude this dissertation by illustrating how this framework can be used not just to autopsy failed policies but to design better ones.

The argument here is that better policy is based on formative, rather than summative measures. This conclusion comes directly from the framework, which forecloses the possibility that there is such a thing as a better summative model. Put simply, summative accountability cannot work in public education because it necessarily satisfies all three conditions of failure. That is, this dissertation did not merely identify a case of poorly designed summative accountability. It provides an illustration of why summative accountability cannot be designed effectively in public education.

Summative accountability in education requires a degree of simplification that inevitably triggers the mechanisms that sow the seeds of failure. Echoing Bonini's paradox, Rowan and Raudenbush's (2016) discussion of distortion implies that creating a metric that is simple enough to be interpreted while capturing the myriad outputs for which teachers are responsible is exceedingly difficult. I would go further and argue that it requires such a degree of simplification that it cannot be done. Summative measures need to be extremely simple. Complex metrics are too difficult to administer and too hard to defend to the subjects of accountability. However, with stakes attached, validity is

critical. The gap, therefore, between model and reality when used for individual summative accountability in public education will always be large and will always be problematic. The metric's limitations will be highly contested ground, triggering resistance and undermining key conditions of successful instruction, including collaboration and the willingness to use data formatively. On the other hand, where a metric is used formatively, the gap between the model and the reality is less problematic. Formative models require less simplification. And the gap between model and reality has fewer implications for the subjects of reform.

Likewise, summative accountability dictates delocalization of expertise because it requires "objectivity" and standardization. Letting the subjects of evaluation participate in the measure of their evaluation flies in the face of bureaucratic logic. Central administrators therefore look to external experts to develop measures. By devaluing local expertise, central administrators lose a valuable asset needed to accurately characterize the problem and plan successful interventions. Likewise, because teachers highly value their discretion and are validated by their role as experts, delocalization triggers resistance and undermines key conditions of success, including control, legitimacy and motivation. Formative reform has none of these limitations.

Finally, summative accountability generally cannot be implemented in a way that reflects full appreciation of context. This is partially a byproduct of the need for simplicity – a metric that fully accounts for context would be highly complex and difficult to administer – and partially a reflection of political reality - applying a summative accountability system differently in different contexts would likely trigger substantial political pushback. Again, this is not an issue for formative reform.

The inevitable next round of education reforms should therefore be built on a formative rather than summative framework. This is consistent with Rowan and Raudenbush's (2016) argument that "…organizations (including schools) will tend to derive more performance benefits from…using such measures as tools in a more information-rich, frequent, and low stakes evaluative context to promote employee learning and professional development." (p. 1205). Stated in the language of the framework developed here, reforms that are entirely formative are more likely to succeed because they do not need to be oversimplified, they can value local expertise and they can be context-dependent. As a result, they are more consistent with an accurate characterization of the problem – lacking knowledge and tools rather than incentive - and teachers' preferences for formative tools. They are less likely to engender resistance and trigger unproductive coping mechanisms. And formative measures are also consistent with the conditions of success, enhancing legitimacy and control, facilitating rather than inhibiting collaboration, and building on rather than diminishing teachers' motivations. Because there is inherent conflict between summative and formative models, however, a formative model cannot succeed unless freed from the limitations of a system governed primarily by accountability.

Bibliography

Abelson, M.A., & Baysinger, B.D. (1984). Optimal and dysfunctional turnover: Toward an organizational level model. The Academy of Management Review, 9(2), 331-341.

Alvesson, M., & Willmott, H. (2002). Identity regulation as organizational control: Producing the appropriate individual. Journal of Management Studies, 39(5), 619-644.

Anderson, C. S. (1982). The search for school climate: A review of the research. Review of educational research, 52(3), 368-420.

Andreatta, S. (1998). Transformation of the Agro-Food Sector: Lessons from the Caribbean. *Human Organization*. *57*(4), 414-429.

Arrow, K. (1974). The Limits of Organization. WW Norton and Company.

Bacolod, M. P. (2007). Do alternative opportunities matter? The role of female labor markets in the decline of teacher quality. The Review of Economics and Statistics, 89(4), 737-751.

Bajaj, A. (2011). The Evolution of the American Teacher Labor Force in the Latter 20th Century: Dimensions of Gender, Race, and Salary.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. J., & Shepard, L. A. (2010). Problems with the use of student test scores to evaluate teachers (EPI Briefing Paper 278). Washington, DC: Economic Policy Institute. Retrieved from http://www.epi.org/page/-/pdf/bp278.pdf

Ball, D. L., & Hill, H. C. (2009). Measuring teaching quality in practice. In D. H. Gitomer (Ed.), Measurement issues and the assessment for teacher quality (pp. 80–98). Thousand Oaks, CA: Sage.

Balu, R. (2010). Examining teacher turnover: The role of school leadership. Politique américaine, (3), 55-79.

Barnes, G., Crowe, E., & Schaefer, B. (2007). The cost of teacher turnover in five school districts. Washington, DC: National Commission on Teaching and America's Future.

Barry, A. (2002). The anti-political economy. Economy and Society, 31(2), 268-284.

Bennett, K. P., & LeCompte, M. D. (1990). How Schools Work: Sociological Analysis of Education. Longman Publishing Group, 95 Church Street, White Plains, NY 10601.

Berry, B., Noblit, G. W., & Hare, R. D. (1985). A qualitative critique of teacher labor market studies. The Urban Review, 17(2), 98-110.

Betebenner, D. (2011). An overview of student growth percentiles. Dover, NH: National Center for the Improvement of Educational Assessment.

Bidwell, C. E. (1965). The school as a formal organization. Handbook of organizations, 972, 1019.

Bidwell, C. E. (2001). Analyzing schools as organizations: Long-term permanence and short-term change. Sociology of Education, 100-114.

Bland, J., Sherer, D., Guha, R., Woodworth, K., Shields, P., Tiffany-Morales, J., & Campbell, A. (2011). The Status of the Teaching Profession 2011. Center for the Future of Teaching and Learning at WestEd.

Blau (1963). Dynamics of Bureaucracy. Chicago: Chicago University Press.
Boal, W. M. (2009). The Effect of Minimum Salaries on Employment of Teachers: A Test of the Monopsony Model. Southern Economic Journal, 75(3), 611.
Boal, W. M., & Ransom, M. R. (1997). Monopsony in the labor market. Journal of Economic Literature, 35(1), 86-112.
Boardman, A. E., Darling-Hammond, L., & Mullin, S. P. (1982). A framework for the analysis of teachers' demand and supply. Economics of Education Review, 2(2), 127-155.
Boe, E. E., & Gilford, D. M. (1992). Teacher supply, demand, and quality: Policy issues, models, and data bases: Proceedings of a conference. National Academies Press.
Boe, E. E., Cook, L. H., & Sunderland, R. J. (2008). Teacher qualifications and turnover: Bivariate associations with various aspects of teacher preparation, induction, mentoring, extra support, professional development, and workload factors for early career teachers in special and general education. Data Analysis Report No. 2008-DAR1). Philadelphia: University of Pennsylvania, Center for Research and Evaluation in Social Policy.
Bonhomme, S., Jolivet, G., & Leuven, E. (2012). Job Characteristics and Labor Turnover: Assessing the Role of Preferences and Opportunities in Teacher Mobility (No. 8841). CEPR Discussion Papers.
Bonini, C.P. (1963) Simulation of information and decision systems in the firm, Englewood Cliffs, N. J.: Prentice-Hall
Borman, G.D., and Dowling, N.M. (2008). Teacher Attrition and Retention: A meta-analytic and narrative review of the research. 78, 367-409.
Borman, G.D., and Dowling, N.M. (2008). Teacher Attrition and Retention: A meta-analytic and narrative review of the research. 78, 367-409.
Bowker, G. C., & Star, S. L. (2000). Sorting things out: Classification and its consequences. The MIT Press.
Boyd, D., Grossman, P., Hammerness, K., Lankford, H., Loeb, S., Ronfeldt, M., & Wyckoff, J. (2012). Recruiting Effective Math Teachers, Evidence From New York City. American Educational Research Journal.
Boyd, D., Grossman, P., Ing, M., Lankford, H., Loeb, S., & Wyckoff, J. (2011). The influence of school administrators on teacher retention decisions. American Educational Research Journal, 48(2), 303-333.
Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2005). How changes in entry requirements alter the teacher workforce and affect student achievement (No. w11844). National Bureau of Economic Research.
Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2008a). Who leaves? Teacher attrition and student achievement (No. w14022). National Bureau of Economic Research
Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2003). Understanding teacher labor markets: Implications for equity. School finance and teacher quality: Exploring the connection, 55-84.
Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2008b). The Impact of Assessment and Accountability on Teacher Recruitment and Retention: Are There Unintended Consequences? Public Finance Review, 36(1), 88-111.

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2010). Analyzing the determinants of the matching of public school teachers to jobs: Disentangling the preferences of teachers and employers. Journal of Labor Economics, 31(1), 83-117.

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

Boyne, G. A. (2003). Sources of public service improvement: A critical review and research agenda. Journal of public administration research and theory,13(3), 367-394.

Bradley, A. (1999). States' uneven teacher supply complicates staffing of schools. Education Week, 18(26), 1.

Braun, H. I. (2005). Using student progress to evaluate teachers: A primer on value-added models. Hopewell, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/PICVAM.pdf

Briggs, D. (2008, November). The goals and uses of value-added models. Paper presented at the workshop of the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, National Research Council, Washington, DC. Retrieved from http://www7.nationalacademies.org/bota/VAM%20Goals%20and%20Uses%20paper%20-%20Briggs.pdf

Brookover, W., Schweitzer, J., Schneider, J., Beady, C., Flood, P. & Wisenbaker, J. (1978). Elementary school social climate and school achievement. American Educational Research Journal, 15(2), 301-318

Bush, G. W. (n.d.). President Signs Landmark No Child Left Behind Education Bill. Retrieved January 4, 2020, from https://georgewbush-whitehouse.archives.gov/news/releases/2002/01/20020108-1.html.

Callahan, R. (1962). Education and the Cult of Efficiency. Chicago: The University of Chicago

Callon, M., & Law, J. (2005). On qualculation, agency, and otherness.Environment and Planning D, 23(5), 717.

Callon, M., & Muniesa, F. (2005). Peripheral Vision Economic Markets as Calculative Collective Devices. Organization studies, 26(8), 1229-1250.

Cannata, M. (2011). The Role of Social Networks in the Teacher Job Search Process. The Elementary School Journal, 111(3), 477-500.

Cannata, M., Demerath, P., Lynch, J., Milner IV, H. R., Peters, A., Davidson, M., & Long, D. (2010). Understanding the teacher job search process: Espoused preferences and preferences in use. Teachers College Record,112(12), 2889-2934.

Carnoy (2002). Does External Accountability Affect Student Outcomes? A Cross-State Analysis. Educational Evaluation and Policy Analysis, 24(4), 305-331.

Carruthers, C. K. (2012). The Qualifications and Classroom Performance of Teachers Moving to Charter Schools. Education Finance and Policy, 7(3), 233-268.

Chetty, R., Friedman, J. N., Rockoff, J. E. (2011). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. NBER Working Paper No. 17699. Cambridge, MA: National Bureau of Economic Research.

Chubb, J. and Moe, T. (1990). Politics, markets, and America's schools. Brookings Institution Press.
Clotfelter, C. T., Glennie, E., Ladd, H. F., & Vigdor, J. L. (2008). Would higher salaries keep teachers in high-poverty schools? Evidence from a policy intervention in North Carolina. Journal of Public Economics, 92, 1352-1370.
Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. Journa l of Human Resources, 41(4), 778-820.
Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2010). Teacher Labor Markets, Segregation and Salary-Based Policies to Combat Inequity across Schools. Society for Research on Educational Effectiveness.
Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2011). Teacher mobility, school segregation, and pay-based policies to level the playing field. Education Finance and Policy, 6(3), 399-438.
Clotfelter, C. T., Ladd, H. F., Vigdor, J. L., & Diaz, R. A. (2004). Do school accountability systems make it more difficult for low-schools to attract and retain high-teachers? Journal of Policy Analysis and Management, 23(2), 251-271.
Clotfelter, C., Ladd, H. F., & Vigdor, J. (2004). Teacher quality and minority achievement gaps. Durham, NC: Terry Sanford Institute of Public Policy. http://www.sanford.duke.edu/research/papers/SAN04-04.pdf
Cohen, D. (2010). Teacher quality: An American educational dilemma. In M. M. Kennedy (Ed.), Teacher assessment and the quest for teacher quality: A handbook (pp. 375–402). San Francisco, CA: John Wiley & Sons.
Cohen, M. D., March, J. G., & Olsen, J. P. (1972). A garbage can model of organizational choice. Administrative science quarterly, 1-25.
Currie, J. (1991). Employment determination in a unionized public-sector labor market: the case of Ontario's school teachers. Journal of Labor Economics, 45-66.
Dahlby, B. G. (1981). Monopsony and the shortage of school teachers in England and Wales, 1948–73. Applied Economics, 13(3), 303-319.
Danielson, C. (2007). Enhancing professional practice: A framework for teaching. Alexandria, VA: ASCD.
Darling-Hammond, L., & Sykes, G. (2003). Wanted, A National Teacher Supply Policy for Education: The Right Way to Meet The" Highly Qualified Teacher" Challenge. Education Policy Analysis Archives, 11, 33.
Daughtrey, A. (2010). Responding to Teacher Attrition. Sanford Journal of Public Policy, 1(1).
Dean, M. (2009). Governmentality: Power and rule in modern society. SAGE Publications Limited.
Dewey, J. (1922a). Education as Engineering. The New Republic, 32(407), 91.
Dewey, J. (1922b). Individuality and Mediocrity. in John Dewey: the middle works, Vol 13.
Dewey, J. (1922c). Individuality, Equality and Superiority. in John Dewey: the middle works, Vol 13.
Dewey, J. (1929). "The Sources of a Science of Education," pp. 1-40 in Dewey, J. The Later Works, Vol. 5: 1929-1930.

Dimaggio, P. and Powell, W. "The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields", American Sociological Review, 48, 147–160.

Doak, D. et. al. (1998). The Statistical Inevitability of Stability-Diversity Relationships. *The American Naturalist*, *151*(3), 264-276.

DOE Archives: Common Core State Standards. (n.d.). Retrieved January 4, 2020, from https://www.state.nj.us/education/archive/sca/

Duncan, A. (2009, July 24). The Race to the Top Begins. Retrieved January 4, 2020, from https://www.ed.gov/news/speeches/race-top-begins.

Economist, the. (2013). "On Your Marks." http://www.economist.com/news/united-states/21589427-states-are-starting-test-teachers-your-marks

Effective Are They, and How Long Do They Stay? Educational Researcher, 41(3), 83-92.

Elmore, R. F. (2000). Building a new structure for school leadership (pp. 1-46). Washington, DC: Albert Shanker Institute.

Etzioni, A. (1964). Modern organisations. Englewood Cliffs: Prentice-Hall.

Falch, T. (2011). Teacher mobility responses to wage changes: Evidence from a quasi-natural experiment. The American Economic Review, 101(3), 460-465. Goldhaber, D., Destler, K., & Player, D. (2010). Teacher labor markets and the perils of using hedonics to estimate compensating differentials in the public sector. Economics of Education Review, 29(1), 1-17.

Fama, E. and French, K. (2004). The CAPM: Theory and Evidence. *Journal of Economic Perspectives, 18*(3), 25-46.

Fischer, F. & Forester, J. (Ed.). (1993). The argumentative turn in policy analysis and planning. Duke University Press.

Flyvberg, B. (2001). Making Social Science Matter: Why Social Inquiry Fails and How it Can Succeed Again. Cambridge: Cambridge University Press.

Foucault, Michel. 1978 [1991]. "Governmentality." In G. Burchell C. Gordon and P. Miller eds. The Foucault Effect: Studies in Governmentality. Chicago: The University of Chicago Press.

Gamoran, A., & Dreeben, R. (1986). Coupling and control in educational organizations. Administrative science quarterly, 612-632.

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G. (2010). Evaluating teachers: The important role of value-added. Washington, DC: Brown Center on Education. Retrieved from http://www.brookings.edu/~/media/research/files/reports/2010/11/17%20evaluating%20teachers/1117_evaluating_teachers.pdf

Goldhaber, D. D., & Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement. Educational Evaluation and Policy Analysis, 22(2), 129–145. Retrieved from http://epa.sagepub.com/content/22/2/129.full.pdf

Goldhaber, D., Gross, B., & Player, D. (2011). Teacher career paths, teacher quality, and persistence in the classroom: Are public schools keeping their best?. Journal of Policy Analysis and Management, 30(1), 57-87.

Gordon, G. L. (1999). Teacher talent and urban schools. Phi Delta Kappan, 81(4), 304-307.

Gordon, R., Kane, T. J., & Staiger, D. O. (2006). Identifying effective teachers using performance on the job (Hamilton Project Discussion Paper 2006-01). Washington, DC: Brookings Institution. Retrieved from http://www.brookings.edu/views/Papers/200604hamilton_1.pdf

Grant, G. (1988). The world we created at Hamilton High. Harvard University Press.

Goener, S., Lietaer, B, Ulanowicz, R. (2009). Quantifying Economic Sustainability: Implications for Free-Enterprise Policy and Practice. *Ecological Economics, 69*, 76-81.

Griffon, D. and Torres-Alruiz, M.D. (2008). On the inherent instability of the monoculture. Organic World Congress. Retrieved from http://orgprints. Org/11931.

Guarino, C. M., Brown, A. B., & Wyse, A. E. (2011). Can districts keep good teachers in the schools that need them most?. Economics of Education Review, 30(5), 962-979.

Guarino, C. M., Santibanez, L., & Daley, G. A. (2006). Teacher recruitment and retention: A review of the recent empirical literature. Review of Educational Research, 76(2), 173-208.

Guthrie, J. W., & Zusman, A. (1982). Teacher supply and demand in mathematics and science. The Phi Delta Kappan, 64(1), 28-33.

Habermas, J. (1970). Towards a Rational Society. Beacon Press.

Habermas, J. (1984). The theory of communicative action, Vol. I. Boston: Beacon.

Hacking, I. (1976). Logic of statistical inference. Cambridge: Cambridge University Press.

Hacking, I. (1991).. "How Should we do the History of Statistics?." In G. Burchell C. Gordon and P. Miller eds. The Foucault Effect: Studies in Governmentality. Chicago: The University of Chicago Press.

Hacking, I. (1986). Making Up People in Reconstructing Individualism, ed., T. Heller et al. Palo Alto: Stanford University Press, 222-236.

Hammersley-Fletcher, L., & Adnett, N. (2009). Empowerment or prescription? Workforce remodelling at the national and school level. Educational Management Administration & Leadership, 37(2), 180-197.

Hanushek, E. A. (2002). Teacher quality. In L. T. Izumi & W. M. Evers (Eds.), Teacher quality (pp. 1–13). Stanford, CA: Hoover Institution Press. Retrieved from http://hanushek.stanford.edu/sites/default/files/publications/Hanushek%202002%20Teacher%20Quality.pdf

Hanushek, E. A. (2010, December). The economic value of higher teacher quality (NBER Working Paper No. 16606). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w16606.pdf?new_window=1

Hanushek, E. A., & Rivkin, S. G. (2010b). The quality and distribution of teachers under the No Child Left Behind Act. The Journal of Economic Perspectives, 24(3), 133-150.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2004a). Why public schools lose teachers. Journal of Human Resources, 39, 326–354.

Hanushek, E., Kain, J., & Rivkin, S. (2004b). Revolving door. Education Next, 77(1), 77–82.

Harris, D. (2011). Value-added methods in education: What every educator needs to know. Cambridge, MA: Harvard Education Press.
Hector, A. and Bagchi, R. (2007). Biodiversity and ecosystem multifunctionality. *Nature, 448,* 188-191.
Henry, G. T., Bastian, K. C., & Smith, A. A. (2012). Scholarships to Recruit the "Best and Brightest" Into Teaching: Who Is Recruited, Where Do They Teach, How
Hensvik, L. (2012). Competition, Wages and Teacher Sorting: Lessons Learned from a Voucher Reform*. The Economic Journal.
Hill, H. C., Kapitula, L, & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. American Educational Research Journal, 48(3), 794–831. Retrieved from http://aer.sagepub.com/content/48/3/794.full.pdf
Hirschman, A. (1970). Exit, Voice and Loyalty: Responses to Decline in Firms, Organizations, and States.. Boston: Harvard University Press.
Historical Context: Overview of New Jersey's Statewide Testing Program. (2016, July). Retrieved January 4, 2020, from https://www.nj.gov/education/assessment/history.shtml.
Ingersoll, R. (1999). The problem of under-qualified teachers in American secondary schools. EducationalResearcher, 28(2), 26–37.
Ingersoll, R. (2001). Teacher turnover and teacher shortages: An organizational analysis. American Educational Research Journal, 38(3), 499-534.
Ingersoll, R. (2002). Holes in the Teacher Supply Bucket. School Administrator, 59(3), 42-43.
Ingersoll, R. (2003). Is there really a teacher shortage. Seattle: Center for the Study of Teaching and Policy.
Ingersoll, R. (2003). Who Controls Teachers Work: Power and Accountability in America's Schools. Cambridge: Harvard University Press.
Ingersoll, R. & May, H. (2012). The magnitude, destinations and determinants of mathematics and science teacher turnover. Educational Evaluation and Policy Analysis, 34(4), 435.
Ingersoll, R., & Perda, D. (2010). Is the supply of mathematics and science teachers sufficient?. American Educational Research Journal, 47(3), 563-594.
Ingersoll, R., Merrill, L., & May, H. (2012a). What are the effects of school accountability on teacher retention? Philadelphia: Consortium for Policy Research in Education, University of Pennsylvania.
Ingersoll, R., Merrill, L., & May, H. (2012b). Retaining teachers: How preparation matters. Educational Leadership, 69(8), 30–34.
Isbell, F., Calcagno, V. Hector, A. et. al. (2011). High Plant Diversity is Needed to Maintain Ecosystem Services. *Nature*, 477, 199-2003.
Jackson, C. K. (2012). School competition and teacher labor markets: Evidence from charter school entry in North Carolina. Journal of Public Economics, 96(5-6), 431-448.
Jacob, B. A. (2007). The challenges of staffing urban schools with effective teachers. The Future of Children, 17(1), 129-153.
Johnson, M. B. (1978). The Effect of Monopsony Power on Teachers' Salaries. State & Local Government Review, 56-61.

Johnson, M. B., & Mack, D. L. (1978). Monopsony in the Market for Public School Teachers: Empirically: A Reply. State & Local Government Review, 112-114.

Johnson, S. M., Berg, J., & Donaldson, M. (2005). Who stays in teaching and why: A review of the literature on teacher retention. Cambridge, MA: Harvard Graduate School of Education. Available at http://assets.aarp.org/www.aarp.org_/articles/NRTA/Harvard_report.pdf

Kay, J., Allen, T., Fraser, R., Luvall, J., Ulanowicz, R. (2008). Can we Use Energy Based Inicators… Introduction to the Workshop Panel on Energy and Environmental Constraints.

Kelley, C. (1997). Teacher compensation and organization. Educational Evaluation and Policy Analysis, 19(1), 15-28.

Koski, W. (2012). Teacher Collective Bargaining, Teacher Quality, and the Teacher Quality Gap: Toward a Policy-Analytic Framework. Harvard Law and Policy Review, 6, 67-90.

Krieg, J. M. (2006). Teacher quality and attrition. Economics of Education Review, 25(1), 13–27.

Ladd, H. F. (2011). Teachers' Perceptions of Their Working Conditions: How Predictive of Planned and Actual Teacher Movement?. Educational Evaluation and Policy Analysis, 33(2), 235-261.

Landon, J. H., & Baird, R. N. (1971). Monopsony in the market for public school teachers. The American Economic Review, 61(5), 966-971.

Lehman, C. and Tilman, D. (2000). Biodiversity, Stability and Productivity in Competitive Communities. *The American Naturalist,* 156(5), 534-552.

Lietaer, B., Ulanowicz, R., Goerner, S. (2010). Is Our Monetary Structure a Systemic Cause for Financial Instability? Evidence and Remedies from Nature. *Journal of Future Studies*, *14*(3), 89-108.

Lin, C. C. (2002). The Shortage of Registered Nurses in Monopsony: A New View from Efficiency Wage and Job-Hour Models. The American Economist, 29-35.

Lindblom, C. E. (1959). The science of" muddling through". Public administration review, 79-88.

Lingard, B. (2011). Policy as numbers: Accounting for educational research. The Australian Educational Researcher, 38(4), 355-382

Link, C. R., & Landon, J. H. (1975). Monopsony and union power in the market for nurses. Southern Economic Journal, 649-659.

Lipsky (1980). Street Level Bureaucracy: Dilemmas of the Individual in Public

Liu, E. and Johnson, S. M. (2006). New Teachers' experiences of Hiring: Late, Rushed, and information-Poor. Educational Administration Quarterly, 42(3), 324-360.

Liu, E., Rosenstein, J., Swann, A., & Khalil, D. (2008). When districts encounter teacher shortages? The challenges of recruiting and retaining math teachers in urban districts. Leadership and Policy in Schools, 7(3), 296–323.

Lortie, D. C. (1969). The balance of control and autonomy in elementary school teaching. The semi-professions and their organization, 1-53.

Lortie, D. C. (1975). Schoolteacher: A Sociological Study. Chicago: University of Chicago Press.

Luizer, J., & Thornton, R. (1986). Concentration in the labor market for public school teachers. Industrial and Labor Relations Review, 573-584.

Mackenzie. (2009). Material Markets: How Economic Agents are Constructed. Oxford: Oxford University Press, 2009). 228 + x pp

Martin, S. M. (2010). Are public school teacher salaries paid compensating wage differentials for student racial and ethnic characteristics?.Education Economics, 18(3), 349-370.

Matland, Richard (1995). "Synthesizing the implementation literature: the ambiguity-conflict model of policy implementation." Journal of Public Administration Research and Theory. 5, 145- 174.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). Evaluating value-added models for teacher accountability. Santa Monica, CA: RAND Corporation. Retrieved from http://www.rand.org/pubs/monographs/2004/RAND_MG158.pdf

Medcalfe, S., & Thornton, R. J. (2006). Monopsony and teachers' salaries in Georgia. Journal of Labor Research, 27(4), 537-554.

Merrifield, J. (1999). Monopsony power in the market for teachers: Why teachers should support market-based education reform. Journal of Labor Research, 20(3), 377-391.

Meyer and Rowan (2006). Institutional Analysis and the Study of Education. Albany: SUNY Press.

Meyer, J. and Rowan, B. (1978). The structure of educational organizations. In Environments and Organization, ed. Marshall W. Meyer, 78-109. San Francisco: Jossey-Bass.

Miller, H. T., & Fox, C. J. (2007). Postmodern public administration. ME Sharpe Inc.

Miller, P. (1992). Accounting and objectivity: the invention of calculating selves and calculable spaces. Annals of scholarship, 9(1/2), 61-86

Miller, P., & Rose, N. (2008). Governing the present: administering economic, social and personal life. Polity Press.

Mintzberg, H. (1979). The structuring of organizations: A synthesis of the research. University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship.

Mitchell, T. (2002). Rule of experts: Egypt, techno-politics, modernity. University of California Press.

Murnane, R. J., & Steele, J. L. (2007). What is the problem? The challenge of providing effective teachers for all children. The Future of Children, 17(1), 15-43.

Naeem, S. (2002). Biodiversity equals instability? *Nature*, 416, 23-24.

National Center for Education Statistics. (2012a). Surveys and Programs: Elementary and Secondary. Available at http://nces.ed.gov/surveys/SurveyGroups.asp?Group=1

National Center for Education Statistics. (2012b). Surveys and Programs: Post-secondary. Available at http://nces.ed.gov/surveys/SurveyGroups.asp?Group=2

National Commission on Teaching and America's Future. (1997). Doing what matters most: Investing in quality teaching. New York: NCTAF.

New Jersey Department of Education. (n.d.). Retrieved December 31, 2019, from https://www.nj.gov/education/finance/rda/dfg.shtml

New Jersey School Performance Reports: (n.d.). Retrieved December 31, 2019, from https://rc.doe.state.nj.us/runreport.aspx?type=district&county=07&district=0680&year=2017-2018.

Niesche, R. (2013). Foucault, counter-conduct and school leadership as a form of political subjectivity. Journal of Educational Administration and History, 45(2), 144-158.

North, D. (1990). Institutions, Institutional Change and Economic Performance. Cambridge: Cambridge University Press.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? Educational Evaluation and Policy Analysis, 26(3), 237-257.

O'Connor, A. (2001). Poverty Knowledge: Social Science, Social Policy and the Poor in Twentieth-Century U.S. History. Hopewell: Hopewell University Press.

Olson, M. (2009). The logic of collective action: public goods and the theory of groups. Boston: Harvard University Press.

Ozga, J. (2009). Governing education through data in England: From regulation to self-evaluation. Journal of Education Policy, 24(2), 149-162.

Perrow, C., Reiss, A. J., & Wilensky, H. L. (1986). Complex organizations: A critical essay (Vol. 3). New York: McGraw-Hill.

Podgursky, M., & Springer, M. (2011). Teacher Compensation Systems in the United States K-12 Public School System. National Tax Journal, 64(1), 165-192.

Podgursky, M., Monroe, R., & Watson, D. (2004). The academic quality of public school teachers: An analysis of entry and exit behavior. Economics of Education Review, 23(5), 507-518.

Powell, W. (1990) Neither Market nor Hierarchy: Network Forms of Organization. Research on Organizational Behavior, 12, 295-336.

Quartz, K., Thomas, A., Anderson, L., Masyn, K., Lyons, K., & Olsen, B. (2008). Careers in motion: A longitudinal retention study of role changing patterns among urban educators. Teachers College Record, 110(1), 218–250.

Randsom, M. R., & Lambson, V. E. (2011). Monopsony, Mobility, and Sex Differences in Pay: Missouri School Teachers. American Economic Review, 101(3), 454-59.

Ransom, M. R., & Sims, D. P. (2010). Estimating the firm's labor supply curve in a "new monopsony" framework: Schoolteachers in Missouri. Journal of Labor Economics, 28(2), 331-355.

Ravitch, D. (2012). The Death and Life of the Great American School System: How Testing and Choice are Undermining Education. New York: Basic Books.

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. Education Finance and Policy, 4(4), 492–519. Retrieved from http://cepa.stanford.edu/sites/default/files/reardon%20raudenbush%20EFP%20VAM%20paper%20resubmission.pdf

Reeves, J. B. (2010). Academic Optimism and Organizational Climate: An Elementary School Effectiveness Test of Two Measures (Doctoral dissertation, The University of Alabama TUSCALOOSA).

Reininger, M. (2012). Hometown Disadvantage? It Depends on Where You're From Teachers' Location Preferences and the Implications for Staffing Schools. Educational Evaluation and Policy Analysis, 34(2), 127-145.

Ritzer, G. (2008). The McDonaldization of Society. Thousand Oaks: Pine Forge Press.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement.Econometrica, 73(2), 417-458.

Rockoff, J. (2004). The impact of individual teachers on student achievement: Evidence from panel data.American Economic Review, 94(2), 247-252

Ronfeldt, M., Loeb, S., Wyckoff, J. (2013). How teacher turnover harms student achievement. American Educational Research Journal, 50(1), 4-36.

Rose, N. (1988). Calculable minds and manageable individuals. History of the Human Sciences, 1(2), 179-200

Rose, N. (1991). Governing by numbers: figuring out democracy. Accounting, Organizations and Society, 16(7), 673-692.

Rose, N., & Miller, P. (1992). Political power beyond the state: problematics of government. British journal of sociology, 173-205.

Rose, N., O'Malley, P., & Valverde, M. (2006). Governmentality. Annu. Rev. Law Soc. Sci., 2, 83-104.

Rothstein, J. (2009). Student sorting and bias in value added estimation: Selection on observables and unobservables (NBER Working Paper No. 14666). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w14666.pdf?new_window=1

Rothstein, J. (2012). Teacher Quality Policy When Supply Matters. National Bureau of Economic Research Working Paper No. 18419. Cambridge, MA: National Bureau of Economic Research.

Rowan, B. & Raudenbush, S. (2016). Teacher Evaluation in American Schools. In D. Gitomer & C. Bell (Eds). *Handbook of Research on Teaching (5th Edition)* (pp 1159 -1216). AERA.

Rumberger, R. W. (1987). The impact of salary differentials on teacher shortages and turnover: The case of mathematics and science teachers. Economics of Education Review, 6(4), 389-399.

Salzman, Hal & Benderly, Beryl Lieff (2019). STEM performance and supply: assessing the evidence for education policy. Journal of Science Education and Technology, 28(1), 9-25

Scheerens, J. (1990). School effectiveness research and the development of process indicators of school functioning. School effectiveness and school improvement, 1(1), 61-80.

Scheerens, J. (1991). Process indicators of school functioning: a selection based on the research literature on school effectiveness. Studies in Educational Evaluation, 17(2-3), 371-403.

Scott (1998) Seeing Like a State. New Haven: Yale University Press.

Shah, M. (2012). The Importance and Benefits of Teacher Collegiality in Schools–A Literature Review. Procedia-Social and Behavioral Sciences, 46, 1242-1246.

Shymansky, J. A., & Aldridge, B. G. (1982). The teacher crisis in secondary school science and mathematics. Educational Leadership, 40(2), 61-62.

Springer, M. G. (2011). New York City's school-wide bonus pay program: Early evidence from a randomized trial. DIANE Publishing.

Sizer, T. (1984). Horace's Compromise: The Dilemma of the American High School. Boston: Houghton Mifflin Company.

Stecher, Brian M., Deborah J. Holtzman, Michael S. Garet, Laura S. Hamilton, John Engberg, Elizabeth D. Steiner, Abby Robyn, Matthew D. Baird, Italo A. Gutierrez, Evan D. Peet, Iliana Brodziak de los Reyes, Kaitlin Fronberg, Gabriel

Weinberger, Gerald Paul Hunter, and Jay Chambers, Improving Teaching Effectiveness: Final Report: The Intensive Partnerships for Effective Teaching Through 2015–2016. Santa Monica, CA: RAND Corporation, 2018. https://www.rand.org/pubs/research_reports/RR2242.html. Also available in print form.

Stephens, John. (1967). The Process of Schooling. New York: Hold, Rinehart, and Winston.

Suspitsyna, T. (2010). Accountability in American education as a rhetoric and a technology of governmentality. Journal of Education Policy, 25(5), 567-586

Taubman , P. (2009). Teaching By Numbers: Deconstructing the Discourse Around Standards and Accountability in Education. New York: Routledge.

Taxpayers Guide to Education Spending. (2019). Retrieved January 4, 2020, from https://www.nj.gov/education/guide/2019/district.shtml.

Taylor, F. (1911). The Principles of Scientific Management. New York.

Taylor, L. L. (2006). Competition and teacher compensation (Vol. 624). Bush School Working Paper.

Teddlie, C. And Reynolds, D. Eds. (2000). The International Handbook of School Effectiveness Research. New York: Falmer Press.

Theobald (2000) Texas Center for Educational Research. (2000). The cost of teacher turnover. Austin: Texas State Board for Educator Certification.

Thornton, R. J. (1978). Monopsony in the Market for Public School Teachers: Really? A Comment. State & Local Government Review, 10(3), 109-112.

Tilman, D., Lehman, C., Bristol, C. (1998). Diversity Stability Relationships: Statistical Inevitability or Ecological Consequence. *The American Naturalist, 151*(3), 277-282.

Tilman, D., Reich, P., Knops, J. (2006). Biodiversity and Ecosystem Stability in a years-long Grassland Experiment. *Nature, 441*, 629-632.

Toch, T., & Rothman, R. (2008). Rush to judgment: Teacher evaluation in public education. Washington, DC: Education Sector. Retrieved from http://www.educationsector.org/usr_doc/RushToJudgment_ES_Jan08.pdf

Townley, B., Cooper, D. J., & Oakes, L. (2003). Performance measures and the rationalization of organizations. Organization Studies, 24(7), 1045-1071

Tyack, D. (1974). The One Best System: A History of American Urban Education. Cambridge.

Ulanowicz, R., Goerner, S., Lietaer, B., Gomez, R. (2007). Quantifying Sustainability: Resilience, Efficiency and the Return of Information Theory. *Ecological Complexity, 6*, 27-36.

United States. National Commission on Excellence in Education. (1983). A nation at risk : the imperative for educational reform : a report to the Nation and the Secretary of Education, United States Department of Education. Washington, D.C. :The Commission : [Supt. of Docs., U.S. G.P.O. distributor].

Vedder, R., & Hall, J. (2000). Private school competition and public school teacher salaries. Journal of Labor Research, 21(1), 162-168.

Villar, A., & Strong, M. (2007). Is mentoring worth the money? A benefit-cost analysis and five-year rate of return of a comprehensive mentoring program for beginning teachers. ERS Spectrum, 25(3), 1–17.

Walberg, H. J. (1991). Productive teaching and instruction: Assessing the knowledge base. In H. C. Waxman & H. J. Walberg (Eds.), Effective teaching: Current research (pp. 33–62). Berkeley, CA: McCutchan.

Waxman, H. C. (1995). Classroom observations of effective teaching. In A. C. Ornstein (Ed.), Teaching: Theory into practice. Needham Heights, MA: Allyn and Bacon.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. Review of Educational Research, 73(1), 89–122. Retrieved from http://web.missouri.edu/~podgurskym/Econ_4345/syl_articles/wayne_youngs_teacher_effects.pdf

Weber, M. (1919) Science as a Vocation, in Gerth, H.H. & Mills, C.W. (eds). From Max Weber: Essays in Sociology. New York: Oxford University Press.

Weber, M. (1924/1968). Bureaucracy and Legitimate Authority in Economy and Society: An Outline of Interpretive Sociology. University of California Press.

Weber, M. (1952[1998]) The Protestant Ethic & The Spirit of Capitalism. Los Angeles: Roxbury Publishing Company.

Weick, K. E. (1976). Educational organizations as loosely coupled systems. Administrative science quarterly, 1-19.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. New York, NY: The New Teacher Project. Retrieved from http://widgeteffect.org/downloads/TheWidgetEffect.pdf

Wilkinson, G. (2006). McSchools for McWorld? Mediating global pressures with a McDonaldizing education policy response. Cambridge Journal of Education,36(1), 81-98.

Williamson, O. E. (1981). The economics of organization: the transaction cost approach. American journal of sociology, 548-577.

Willmott, H. (1993). Strength is Ignorance; Slavery is Freedom: Managing Culture in Modern Organizations. Journal of management studies, 30(4), 515-552

Wilson, J. Q. (1989). Bureaucracy: What Government Agencies Do and Why They Do it. New York: Basic Books.

Wright, S., Horn, S., & Sanders, W. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. Journal of Personnel Evaluation in Education, 11, 57–67. Retrieved from http://www.sas.com/govedu/edu/teacher_eval.pdf

Appendix

Teacher Interview Protocol

1. Introduction:
    a. What do you teach?
    b. How long have you been teaching? Overall? This school? This grade? Tenured?
    c. What made you decide to become a teacher? Elementary school? Here?
2. What do you consider to be your primary job/role/duties as a teacher?
3. Tell me about the teacher evaluation system.
    a. What are the main elements of the system? What are you evaluated on? Where do tests come into play? How often are you observed? What observation rubric is used?
    b. Are there expectations that you teach a certain way? Use certain methods?
    c. What type of teacher do you think fares/would fare best? What kind of teacher do you think they are looking for?
    d. Have you or will you receive any pedagogical training?
    e. What are the consequences for you of a poor rating? Of a good rating?
    f. How different is it then when you first started teaching?
4. Tell me about the tests your students take? Are they common core aligned? How much time do you spend administering tests? Preparing for tests? More time on tested subjects?
    a. Different from the past?

5. Have recent changes affected how you feel about your job? Do you consider your role differently now?
6. Have recent changes affected how you do your job?
    a. Why/why not? Incentives directly affected?
    b. What do you do differently?
        i. Content?
        ii. Methods?
        iii. More or less time with low-performing students? High-performing?
    c. Creativity affected?
7. Has your relationship with your principal/vp changed?
8. Did your teacher preparation program prepare you in a way that's consistent with the demands of the changes?
9. How do you feel about the changes?
    a. Necessary? Well implemented? Appropriate?
    b. Good feedback? Will it help you be a better teacher?
    c. Would you have entered teaching if you knew? Have recent changes affected whether or not or how long you will stay? Be specific about which changes.
    d. What type of system would you prefer/recommend?

Principal Interview Protocol

1. How long have you been a principal? This district? School?
2. Were you a teacher first? For how long? What was your admin training?
3. Can you describe what if any changes have taken place as a result of new requirements like TeachNJ, PARCC, CCSSs?
4. Has your role changed as a result of the implementation of PARCC, TeachNJ, CCSSs adoption?
    a. More focus on instruction? More time in the classroom?
    b. Do you take steps to directly influence what teachers teach and how?
5. How do you think teachers are responding to the changes? Any increases in turnover?
6. How do you feel about the changes?

Framework for Teaching

Domain 1 Planning & Preparation	Domain 2 Classroom Environment	Domain 3 Instruction	Domain 4 Professional Responsibilities
A. Demonstrating Knowledge of Content and Pedagogy i. Knowledge of Content and the Structure of the Discipline ii. Knowledge of Prerequisite Relationships iii. Knowledge of Content-Related Pedagogy B. Demonstrating Knowledge of Students i. Knowledge of Child and Adolescent Development ii. Knowledge of the Learning Process iii. Knowledge of Students' Skills, Knowledge, and Language Proficiency iv. Knowledge of Students' Interests and Cultural Heritage v. Knowledge of Students' Special Needs C. Selecting Instructional Outcomes i. Value, Sequence, and Alignment ii. Clarity iii. Balance iv. Suitability for Diverse Learners D. Demonstrating Knowledge of Resources i. Resources for Classroom Use ii. Resources to Extend Content Knowledge and Pedagogy iii. Resources for Students E. Designing Coherent Instruction i. Learning Activities ii. Instructional Materials and Resources iii. Instructional Groups iv. Lesson and Unit Structure F. Designing Student Assessment i. Congruence with Instructional Outcomes ii. Criteria and Standards iii. Design of Formative Assessments iv. Use for Planning	A. Creating an Environment of Respect and Rapport i. Teacher Interaction with Students ii. Student Interactions with One Another B. Establishing a Culture for Learning i. Importance of the Content ii. Expectations for Learning and Achievement iii. Student Pride in Work C. Managing Classroom Procedures i. Management of Instructional Groups ii. Management of Transitions iii. Management of Materials and Supplies iv. Performance of Non-Instructional Duties v. Supervision of Volunteers and Paraprofessionals D. Managing Student Behavior i. Expectations ii. Monitoring of Student Behavior iii. Response to Student Misbehavior E. Organizing Physical Space i. Safety and Accessibility ii. Arrangement of Furniture and Use of Physical Resources	A. Communicating with Students i. Expectations for Learning ii. Directions and Procedures iii. Explanation of Content iv. Use of Oral and Written Language B. Using Questioning and Discussion Techniques i. Quality of Questions ii. Discussion Techniques iii. Student Participation C. Engaging Students in Learning i. Activities and Assignments ii. Grouping of Students iii. Instructional Materials and Resources iv. Structure and Pacing D. Using Assessment in Instruction i. Assessment Criteria ii. Monitoring of Student Learning iii. Feedback to Students iv. Student Self-Assessment and Monitoring of Progress E. Demonstrating Flexibility and Responsiveness i. Lesson Adjustment ii. Response to Students iii. Persistence	A. Reflecting on Teaching i. Accuracy ii. Use in Future Teaching B. Maintaining Accurate Records i. Student Completion of Assignments ii. Student Progress in Learning iii. Non-Instructional Records C. Communicating with Families i. Information About the Instructional Program ii. Information About Individual Students iii. Engagement of Families in the Instructional Program D. Participating in a Professional Community i. Relationships with Colleagues ii. Involvement in a Culture of Professional Inquiry iii. Service to the School iv. Participation in School and District Projects E. Growing and Developing Professionally i. Enhancement of Content Knowledge and Pedagogical Skill ii. Receptivity to Feedback from Colleagues iii. Service to the Profession F. Demonstrating Professionalism i. Integrity and Ethical Conduct ii. Service to Students iii. Advocacy iv. Decision Making Compliance with School and District Regulations