Teacher Evaluation Systems: How Teachers and Teacher Quality are (re)Defined by Market-Based Discourses

by

Jessica Holloway-Libell

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved October 2014 to the
Graduate Supervisory Committee:

Audrey Amrein-Beardsley, Co-Chair
Kate T. Anderson, Co-Chair
David C. Berliner

ARIZONA STATE UNIVERSITY

December 2014

ABSTRACT

Teacher evaluation policies have recently shifted in the United States. For the first time in history, many states, districts, and administrators are now required to evaluate teachers by methods that are up to 50% based on their "value-added," as demonstrated at the classroom-level by growth on student achievement data over time. Other related instruments and methods, such as classroom observations and rubrics, have also become common practices in teacher evaluation systems. Such methods are consistent with the neoliberal discourse that has dominated the social and political sphere for the past three decades. Employing a discourse analytic approach that called upon a governmentality framework, the author used a complementary approach to understand how contemporary teacher evaluation polices, practices, and instruments work to discursively (re)define teachers and teacher quality in terms of their market value.

For the first part of the analysis, the author collected and analyzed documents and field notes related to the teacher evaluation system at one urban middle school. The analysis included official policy documents, official White House speeches and press releases, evaluation system promotional materials, evaluator training materials, and the like. For the second part of the analysis, she interviewed teachers and their evaluators at the local middle school in order to understand how the participants had embodied the market-based discourse to define themselves as teachers and qualify their practice, quality, and worth accordingly.

The findings of the study suggest that teacher evaluation policies, practices, and instruments make possible a variety of techniques, such as numericization, hierarchical surveillance, normalizing judgments, and audit, in order to first make teachers objects of knowledge and then act upon that knowledge to manage teachers' conduct. The author also

found that teachers and their evaluators have taken up this discourse in order to think about and act upon themselves as responsibilized subjects. Ultimately, the author argues that while much of the attention related to teacher evaluations has focused on the instruments used to measure the construct of teacher quality, that teacher evaluation instruments work in a mutually constitutive ways to discursively shape the construct of teacher quality.

DEDICATION

I dedicate this dissertation to all of my past, present, and future teachers—

both traditional and not.

ACKNOWLEDGMENTS

While this project might have my name on it, it by no means was made possible by me alone. Having the support of the many friends, family members, and mentors has meant the difference between finishing this dissertation and not. For that reason, I would like to thank those people now.

First, I would like to thank my mentor, advisor, and role model, Dr. Audrey Amrein-Beardsley. After adopting me during the first year of the program, you have opened doors and opened my mind. You have pushed me, taught me, and most importantly, supported me. I will never have all the words to express my deepest appreciation that I have for you.

I would also like to thank my other mentor and role model, Dr. Kate Anderson. You always find a way to make me feel just a little less crazy when I am lost. I appreciate the way in which you always entertain my ideas and thoughts, regardless of how irrational (or sometimes right!) they may be. You have encouraged me to be better, stronger, and smarter. Thank you for that.

Other professors have served in various supportive ways that I would also like to acknowledge. Dr. David Berliner, to have had the opportunity to work with you has been a true honor. Dr. Jeanne Powers, as the first person I met in the program, thank you for your unwavering support from day one. Dr. Gustavo Fischman, thank you for being tough on me and ultimately making me stronger. Dr. Sarah Polasky, thank you for your continued support and your help in connecting me with my dissertation site. And, Dr. Jory Brass, thank you for stepping into my life at just the right time. You will never know how much I appreciate your support and friendship.

Also, thank you to the teachers, principals, and district administrators at my dissertation site. You welcomed me into your school, your classrooms, and your lives. It was an honor to have been a part of your team for the past few years. Thank you.

The friends and colleagues I have made on this journey have also contributed to a successful and (usually) joyful ride! Dr. Noelle Paufler, thank you for your ongoing friendship and support. Dr. Clarin Collins, thank you for being a constant inspiration and for encouraging me the find the right light and write there. Taucia Gonzalez, thank you for your patience, wisdom, guidance, inspiration, and friendship. Jesus Cisneros, thank you for being you—I treasure your friendship more than you know. Laura Gomez, I am quite certain that I would not be the person I am today without you. You somehow ground me yet also encourage me to think more deeply. Though we may be parting, I know that our friendship will only grow stronger from here.

Lastly, to my family, thank you. To my parents, you have never allowed me to see boundaries, regardless of the circumstances; for that reason I have always lived in a world of endless possibilities. To Tyler, I will never be able to fully thank you for your undying support. You make me laugh and let me cry. There is no way I could have done this without you. Thank you for letting me live out my dream.

TABLE OF CONTENTS

Page

APPENDIX Page

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER 1

Introduction and Overview

We currently live in an era dominated by the need to count, measure, compare, and evaluate nearly every aspect of society. Nikolas Rose (1999) wrote, "the apparent objectivity of numbers, and of those who fabricate and manipulate them, helps configure the respective boundaries of the political and the technical. Numbers are part of the techniques of objectivity that establish what it is for a decision to be 'disinterested'" (p. 199). Numbers provide us a way of making objective, rational comparisons and decisions, especially as they relate to resource allocations and social programs. Public education has been no exception, and teachers, specifically, have most recently been subjected to such practices.

The conceptualization of social matters as numbers is not new, as the social sciences, in particular, have attempted to use statistics for more than two centuries to understand human behavior and other social phenomena. Statistics allowed for populations to be understood as objects of knowledge and thus acted upon and governed. The way in which such populations have been governed has changed over time, in accordance to what Foucault has termed "governmentality" (1980; 1991), or the rationality of governance. Here, "governance" does not simply refer to the political or official bureaucratic sense of the word, but rather a mode of management that can relate to the management of subjects or the self. Over time, these rationalities of governance have changed; in this particular sociohistoric moment, neoliberalism is the dominant governmentality (Lemke, 2002). As such, social matters and public institutions are reconfigured as market-based entities that are made sense of, valued, and acted upon in

terms of their market value. Education has been one such institution.

Teacher evaluation practices, in particular, have shifted accordingly, undergirded by the argument that America's public school teachers are lacking in quality. Federal financial incentive programs such as Race to the Top (RttT), the Teacher Incentive Fund (TIF) grants program, and Elementary and Secondary Education Act (ESEA) waivers (i.e., waivers to exempt schools from meeting requirements previously established by No Child Left Behind), have provoked systematic changes by incentivizing states, and thus school districts, to develop methods for identifying, and in some cases firing, America's purportedly subpar teachers. Accordingly, for the first time in history, many states, districts, and administrators, are now required to evaluate teachers by methods that are up to 50% based on their "value-added," as demonstrated at the classroom-level by growth on student achievement data over time (RttT, 2011).

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

Counter to these claims, opponents, including teachers, educational researchers, and grassroots education advocates, have responded in public and academic ways. For

example, teacher evaluations were at the forefront of the 2012 Chicago Teachers Strike due to the heavy reliance evaluations were to have on student achievement data (Tareen, 2012). Diane Ravitch, an education scholar and blogger about educational issues, has devoted nearly 500 posts to the topic of teacher evaluations alone (see http://dianeravitch.net/category/teacher-evaluations/). Additionally, critics of the Chetty et al. (2011) study indicated that increased earning potential resulting from effective teachers broke down to less than $20 per week per student (Baker, 2012), that the study was based on data prior to NCLB (Winerup, 2012), and that the researchers contradicted themselves in their findings, thus invalidating their claims (Adler, 2013). In all, opponents have argued that the current methods of measuring teacher effectiveness based on student growth are vastly flawed, primarily in terms of reliability, validity, bias, and fairness (Baker, Oluwole, & Green, 2013; Berliner, 2014; Hill, Kapitula, & Umland, 2011; Papay, 2010).

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

Interestingly, while the debate about teacher quality (and how to measure it) has grabbed the attention of the public, it does not follow typical partisan boundaries like

most contemporary social issues (e.g., immigration, health care, etc.). Instead, proponents of the policies include legislatures from both sides of the aisle, while opponents include a range from progressives to tea party affiliates (e.g., the Badass Teachers Association, or BATs, are a heterogeneous group of grassroots organizers whose mission is to fight against policies such as VAM-based teacher evaluations). What this does not mean, however, is that the debate is any less contentious, or any less binary. As the titles of recent books about the debate, such as Ravitch's (2013) *Reign of Error: The Hoax of the Privatization Movement and the Danger to America's Public Schools* and Berliner and Glass's (2014) *50 Myths and Lies that Threaten America's Public Schools*, suggest, there is a clear dichotomy of winners (e.g., politicians, financiers, etc.) and losers (e.g., teachers, students, etc.).

While I do not grapple with the notion that there are those who benefit from the system and also those who suffer, I do want to propose an alternative perspective and offer a critique that may not be as black and white. To do so, I will use Rose's (1999) conceptualization of the relationship between numbers and society, situated within a governmentality (Foucault, 1991) framework, to argue that 1) evaluation systems that are designed to measure teaching numerically are consistent with the neoliberal discourse of the present sociohistoric moment, and 2) that such systems work to (re)define teachers and teaching quality, while simultaneously producing particular types of teachers who behave in desired, governable ways.

Globalization, Standardization, and Education

Globalization is defined here broadly as a global-based market where economies, products, cultures, people, ideas, and so forth are no longer confined to traditional nation-

states, but are rather part of a global exchange, or what Robertson (1990, p. 8) simply called the "world-as-a-whole." Such a reformation of global relationships has changed the way we think about competition, which has also repositioned education in terms of its function in society. Since the mid 1980s—after the release of *A Nation at Risk* (U.S. Department of Education, 1983)—public education has been positioned as an economic-based mechanism for individuals and the nation to succeed in an ever-changing, global society (Holloway-Libell & Collins, 2014). In light of such a change, concepts such as accountability, rigor, high-stakes testing, value-added, and, most recently, college and career readiness, have become part of the common vernacular in discussions about education, which have manifested in the form of policies and practices regarding schools, students, and teachers.

Of particular interest to this study is teachers and how in this era of globalization, teachers and teacher quality have been problematized and thus (re)defined in terms of their market value. However, to discuss this issue, I must first discuss the presence of numericization (Rose, 1999) in education. Numericization, or the process of translating abstract ideas (e.g., intelligence, learning, teaching, etc.) into numbers, is not new in education (or other fields). The social sciences have relied heavily on statistics to make sense of our social world, including the field of education. Given the changing global landscape as discussed, an increased reliance on numerical representations of education-related matters has surfaced. Related, a neoliberal discourse that requires capital (e.g., economic, social, human, etc.) to be made measurable, comparable, and evaluate-able has grown to dominate the way we make sense of society since the 1980s (Peters, 1996). As such, various techniques have been used to make various aspects of education into

objects of knowledge that can then be measured, compared, and evaluated. Standards, for example, allow for normalized judgments to be made (Foucault, 1977), which can also make possible measurement and comparison.

Further, standards (e.g., curriculum, teaching, etc.) are a way in which students, teachers, and other school subjects can compare *themselves* against that which has been accepted as normal, allowing them (and encouraging them) to adjust their behaviors accordingly. Policies, practices, and instruments have been developed and implemented in order to make sure that such school subjects are behaving in the desired (or standardized) way. This numericization thus makes possible the techniques of hierarchical surveillance, or the constant visibility of teachers, and examination, or the observation and judgment of teachers (Foucault, 1977; Rose, 1999)—these two technologies work to discipline teachers and produce self-disciplined teachers. Graham and Neu (2004) wrote the following on Foucault's (1984) concept of examination:

> Foucault suggests that examinations impose on examinees a compulsory visibility. Through examinations, attributes of the examinees are made visible, thereby enmeshing the examinees in particular relations of power. Furthermore, the permanent accumulation of these documentary traces in government files and databases introduces individuality into the field of documentation and 'constructs' the examinee as a 'case'. This case 'is the individual as he [sic] may be described, judged, measured, compared with others, in his very individuality; and it is also the individual who has to be trained or corrected, classified, normalized, excluded, etc.' (p. 300).

Teachers, specifically, have become subjected to various forms of evaluation that

work to numericize aspects of their teaching practice and quality. Observation rubrics and value-added models (VAMs) are two of the most prominent instruments for doing so. Rubrics are typically comprised of a set of standards that are accompanied by some numerical rating system to judge the level of performance as it relates to the respective standard. Rubrics are commonly used for classroom observations and other professional responsibilities. VAMs are statistical tools that are designed to measure student growth by comparing student achievement scores on standardized tests over time and then attributing that growth to teachers. Both of these instruments attempt to capture components of teaching quality and make it visible, measurable, comparable, and evaluate-able.

As such, teachers (and teaching) can be thought about as part of a production function where teachers produce a product (knowledge) that is consumed by students. The teachers, then, are thought to either "add" or "detract" value from such a function. This shift in the conceptualization of teachers and education has been in steady progress since the mid 1970s when economist, Erik Hanushek (1971) argued for a better understanding of the inputs and outputs associated with education. Since then, political leaders, such as President Bush, and most recently, President Obama and Secretary of Education Duncan, have adopted this discourse and have framed their campaign for RttT and other federal policies regarding teacher quality and evaluation systems as a response to the economic health of the country. In doing so, teachers and teacher quality have been (re)defined in terms of their overall market value. Peters (1996) argued:

> There is perhaps no better example of the extension of the market to new areas of social life than the field of education. In particular, it is clear that under principles

> of neoliberalism education has been discursively restructured according to the logic of the market. Education, in this model, is treated no differently from any other service or commodity, (p. 81).

The Problem and Purpose

When issues are numericized, the conversations are no longer about the issue itself, but rather about the methods used to measure it—it becomes a technical debate rather than an ideological one (Prewitt, 1987; Rose, 1999; Starr, 1987). As such, a vast majority of the research on teacher evaluation policies, practices, and instruments has focused on the technical properties of such issues, such as the reliability, validity, bias, and fairness of VAMs (Baker et al., 2013; Berliner, 2013; Hill et al., 2011; Papay, 2010), or the intended and unintended effects of such policies (Amrein-Beardsley & Collins, 2012; Collins, 2012). Some scholars have taken on the ideological issue, attempting to reveal an agenda behind the current education reform movement, including teacher evaluations (Amrein-Beardsley, 2014; Berliner & Glass, 2014; Ravitch, 2013). These exposés have shed light on some of the political and financial interests that have likely shaped the current policy landscape that dictates education matters. Another thing this work has in common is that it frames the issue as a dichotomous matter made up of wrongdoers (e.g., financial investors, politicians, conservative think tanks, etc.) and victims (e.g., schools, teachers, students, the public, etc.). Again, while I do not deny the position that there are some people who gain from the system and some who lose, what I do propose is that we take on the issue from a different vantage point to add a missing, but complementary, critique.

Instead of focusing on the intentions behind the policies, or the reliability, validity,

bias, or fairness of the instruments used to carry out the policies, I intend to shift the focus by posing the question: how do the policies and instruments work to problematize teachers and teacher quality in particular ways? In other words, instead of questioning how well or fairly an instrument measures the construct of teacher quality, I am interested in how the instrument works to define the very construct that it intends to measure. Additionally, I am interested in how this process affects the way in which teachers take up and embody the (re)defined construct as it relates to their teaching, quality, and worth. Thus, the purpose of this study is to present a discursive analysis that challenges the way we think about the function of evaluation policies, practices, and instruments. My ultimate goal for this study is to break away from the common frameworks for which we think about teachers and teacher quality in hopes of opening space for new possibilities.

Research Questions

In order to accomplish this purpose, I started with five guiding questions. However, the analysis was an iterative and reflexive process that included renegotiating and re-theorizing along the way, as informed by data collection, analysis, and the literature. The questions below remained at the core of the study:

1. How are teachers positioned as the problem within contemporary teacher evaluation policies and policy discussions?
2. How do teacher evaluation policies, practices, and instruments problematize teachers and teacher quality in particular ways? Or, how are teachers and teacher quality (re)defined by evaluation policies, practices, and instruments?
3. How do teachers embody market-based discourses in talking about and defining themselves, as well as their practice, quality, and worth?

Overview of the Dissertation

In Chapter 2, I provide a review of the current literature on evaluation policies, practices, and instruments. I start with an historical account of how the policies (e.g., Race to the Top, Teacher Incentive Fund grants) came into being by looking at 30 years of education reform trends. Then I look specifically at the instruments used to carry out the policies—value-added models and observation rubrics. I discuss the empirical research on such tools, with a specific focus on the reliability, validity, bias, and outcomes. In the second half of the chapter, I lay out the current teacher evaluation policy landscape as it directly relates to the local context of the study, Desert Middle School (pseudonym). Here I cover each level of policy—federal, state, and local. I recently published a similar, but different, version of Chapter 2 titled "VAM-Based teacher evaluation policies: Ideological foundations, policy mechanisms, and implications," (Holloway-Libell & Collins, 2014). This chapter also contains parts of this (2014) article.

In Chapter 3, I provide the theoretical and methodological framework within which I developed and conducted the study. I start with a discussion of my theoretical transition from a critical structural approach to a poststructural approach. Then I detail how I have defined and operationalized discourse for the purpose of this study, followed by an explanation of Foucault's (1991b) governmentality framework and related concepts. I also discuss neoliberalism as a governing strategy and its relevance in the shaping of the evaluation methods in question. In the second half of the chapter, I link this framework to the shaping of a methodological approach I used to answer the research questions. Utilizing discourse analysis, I developed a two-way analytic approach that allowed me to look at both the problematization of teachers and teacher quality in light of market-based

discourses, as well as how teachers have embodied this way of talking about themselves, and their practice and quality. For the first approach, which I used to answer questions related to the problematization of teachers and teacher quality, I (1) collected and analyzed official documents related to the policies, practices, and instruments that were relevant to the local context of the study; and (2) attended the official evaluator training course where I was able to collect and subsequently analyze field notes and evaluator training materials related to the specific evaluation methods, practices, and instruments of the study's school site. For the second approach, which I used to answer questions related to the teachers' embodiment of market-based discourses, I interviewed teachers and their evaluators at one Arizona middle school. I used these data to make sense of how teachers and their evaluators have taken up the discourse to talk about and act upon themselves accordingly.

In Chapter 4, I present the results of the first approach. First, I collected and analyzed policy-related literature, including official policy documents, political speech transcripts, promotional materials, and the like. I also attended the 35-hour TAP (i.e., the comprehensive teacher evaluation system at Desert Middle School) evaluator certification course where I collected field notes and training materials. For the first part of the analysis, I sought to understand how teachers had been positioned as a problem within education policy discussions. I found that teachers have been positioned as risky subjects and in need of being managed by policies, practices, and instruments. The teachers have been presented as part of a greater market-based model of education, which is consistent with a neoliberal discourse. Accordingly, teacher evaluation technologies have been developed (and adopted at the local level) to manage the risk related to

teachers, as well as produce teachers who manage themselves.

In Chapter 5, I present the results of the second approach where I attempt to link the teacher evaluation techniques to the way in which teachers have begun to talk about themselves, their practice, quality, and worth. To do so, I interviewed a group of teachers and their evaluators at one middle school. In this analysis I demonstrate how teachers have embodied the neoliberal discourse, and in so doing, have begun to define themselves and qualify their practice, quality, and worth in terms of market value. Similarly, they have subjected themselves to various techniques of governance, while denouncing other teachers who have chosen not to participate. This justification rests on a binary that the teachers have constructed about what it means to be an acceptable versus an unacceptable Desert Middle School teacher.

In Chapter 6, I bring the two approaches together and link the evaluation techniques discussed in Chapter 4 to the way in which the teachers have defined themselves and their quality in Chapter 5, drawing conclusions from the findings to answer the research questions. I also provide brief sections on the challenges I faced and lessons I learned during the course of the study. Finally, I provide implications for policy, practice, and further research. Perhaps most importantly, I argue that, as policy analysts, we should consider alternative ways of thinking about how policy works recursively to not only solve problems, but also constitute problems.

CHAPTER 2

Policies, Practices, and Instruments—A Review of the Literature

In this chapter, I take a look at the research on teacher evaluation policies, practices, and instruments. I start with an overview of the current policy context, and then I provide an historical account of how such methods have come to be. This is followed by a review of the literature on the two core instruments used in teacher evaluations—value-added models (VAMs) and observation rubrics. I discuss the empirical research related to the reliability, validity, and bias of such tools, as well as the associated outcomes and recommendations. In the second half of the chapter, I lay out the current policy landscape that has shaped the evaluation system at Desert Middle School. In this section, I cover the federal, state, and local policy levels.

A Move Towards Market-Based Teacher Evaluation Systems

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

A decade later, the Civil Rights Act of 1964 required a national report on the

equal educational opportunities available for all individuals, catalyzing an accountability movement in the U.S. public education system. Sociologist James Coleman (1966) found inequities across schools including class sizes, student achievement levels, school quality, school resources, and teacher quality as measured by the education levels and training of teachers. In his influential Coleman Report, he reported that teacher quality had the greatest impact on student achievement compared to all other school-related factors. The Coleman Report first introduced the impact of school inputs on student achievement and argued that variation in teacher quality had a cumulative effect on students as they progressed through school (Hanushek, 1979).

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

To further investigate the relationship between inputs and outputs, Hanushek (1970) conducted a study in a school district in southern California where he tracked students from first through third grade to examine the relationship between school system

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

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

Hanushek's econometric model was one of the first *value-added* models derived

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

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

The explicit policy impact of *A Nation at Risk* was first realized in the 1990s with

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

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

than NCLB.

Race to the Top. Simultaneously, The New Teacher Project released a report called "The Widget Effect," purporting that, once again, America's public school children were in danger (Weisberg, Sexton, Mulhern & Keeling, 2009); this time faulty teacher evaluations were to blame for U.S. student achievement lagging behind in the global economy. The authors condemned school administrators' inability to distinguish good teachers from bad, while likening teachers to "widgets," or simply "interchangeable parts," (Weisburg et al., 2009, p. 4). They blamed inadequate teacher evaluation systems, which by their claims rated, on average, 99% of all teachers as effective and 1% the inverse. It seemed the country faced yet another "manufactured crisis" (Berliner & Biddle, 1995), but akin to the influence of *A Nation at Risk*, this new report coupled with similar studies, had significant political influence (Corcoran, 2010; Goldhaber & Hansen, 2010; Hanushek, 2011). Thus, the race was on for a more objective, discerning teacher evaluation system that could "properly" identify effective, average, and ineffective teachers.

RttT (2011) and other post-NCLB policy initiatives, such as the aforementioned TIF grants program, adopted the "widget effect" ideology that schools were failing, teachers were to blame, and that by holding teachers accountable (i.e., punishing bad teachers and rewarding good teachers), teachers would work harder and teach better. Popular media sources, including, for example, news journalists, documentarians, and film producers who had subscribed and/or contributed to these propaganda, helped disseminate, reaffirm, and perpetuate these ideological perspectives in the greater public domain by means of emotive petitions and appeals. For example,

some filmmakers used full-length movies, such as *Waiting for Superman* and *Won't Back Down*, to depict teachers and teachers unions as the epitome of the education "crisis," (Dalton, 2013).

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

The Mechanisms of Market-Based Teacher Evaluations

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

instruction and students' learning.

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

While VAMs are one of the most popular methods of measuring teacher quality via student test scores, it is not the only way. Student Growth Percentile (SGP) models are also quite popular. These two models function similarly in that they both attempt to attribute student achievement to teachers; however they differ in the way they attempt to accomplish this goal. VAMs, specifically, are multivariate models that attempt to isolate teacher effects by statistically controlling for other variables that might affect student achievement, such as gender, race, socioeconomic status, language proficiency level, special need status, and any other data that are available. SGPs, on the other hand, are normative models that do not statistically control for other variables, but rather use a student's previous test score(s) to make predictions about the student's expected growth (as determined based on peers who test similar to them) (Betebenner, 2011; Castellano & Ho, 2013).

Though variations of VAMs exist with different inputs or variables and controls included in the models, the output is always measured by student growth on some type of

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

Reliability and VAMs. In terms of VAMs, reliability refers to the likelihood of a teacher being correctly identified as either adding or detracting value from students' learning. A key marker of reliability would be the consistency of teacher-level value-added scores from one year to the next. Of primary concern here is that evidence of reliability, or stability, is weak to moderate at best, with most value-added researchers yielding time-series correlations within the range of $0.3 \leq r \leq 0.4$ (McCaffrey, Sass, Lockwood, & Mihaly, 2009; Kane & Staiger, 2012; Lockwood & McCaffrey, 2009; Newton, Darling-Hammond, Haertel, & Thomas, 2010), while some correlations are as low as $r = 0$ (Linn & Haug, 2002) or as high as $r = 0.6$ (Kersting, Chen, & Stigler, 2013). This instability can mean one of two things—either a majority of teachers' effectiveness truly fluctuates from one year to the next, or, more likely, there is a reliability problem with the models, which results in the misclassification of teachers. The question remains,

how much error is too much error, especially given the often high stakes attached to such classifications?

Validity and VAMs. Researchers have also questioned the evidence of VAMs' validity, or the model's ability to capture the construct of teacher quality, arguing that many model types cannot fully account for the impact of uncontrollable factors (e.g., other teachers' effects, students' peer effects, summer gains/losses, outside-of-school variable effects, missing data) on yielding valid value-added estimates from which valid inferences can be made (Amrein-Beardsley, 2008; Capitol Hill Briefing, 2011; Ishii & Rivkin, 2009; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Scherrer, 2011).

Additionally, there are issues with criterion-related evidence of validity, which refers to the extent to which value-added scores align with other evaluative measures (Bill & Melinda Gates Foundation, 2013; Papay, 2010), and construct-related evidence of validity, which refers to the extent to which value-added scores actually measure the construct of interest, teaching effectiveness (Capitol Hill Briefing, 2011; Newton et al., 2010; Rothstein, 2009; 2010). First, there is a lack of statistical correlation between value-added estimates and other indicators of teacher quality, such as principal observations or teaching awards (Amrein-Beardsley & Collins, 2012; Collins, 2012). There is also a misalignment between value-added estimates derived from different tests meant to measure the same thing and administered at the same time. This misalignment is approximately $0.37 \leq r \geq 0.5$ for reading/language arts and $0.22 \leq r \geq .59$ for mathematics (Bill & Melinda Gates Foundation, 2010; Corcoran, Jennings, & Beveridge, 2011). There are also concerns when comparing estimates derived from criterion-referenced

assessments to norm-referenced assessments, meaning the scores serve different purposes and do not fairly lend to comparison (Amrein-Beardsley & Collins, 2012).

Bias and VAMs. Yet another point of contention with VAMs is bias (Hill et al., 2011; Newton et al., 2010; Rothstein, 2009), or the extent to which exogenous variables influence teachers' value-added scores and/or their capacities to demonstrate growth (Linn & Haug, 2002; Wright et al., 1997). For example, teachers of students who typically score in the 99th percentile have a difficult time demonstrating growth because there is no room to grow – a phenomenon sometimes called the ceiling effect. Rothstein (2009) argued that there might, theoretically, be ways of mitigating the bias inherent in VAM estimates, such as including more years of data. But he contends that doing so is not always realistic in that there are issues with missing data, as well as the problem that some grades levels can only have one year of data (e.g., third grade). Another recommendation for statistically dealing with such bias would be to randomly assign students and teachers to classrooms (Raudenbush, 2004). However, again, the practical implications of this is very limited in that, principals rarely randomly assign students and teachers to classrooms because they find value in placing students with teachers based on students' needs (Bill & Melinda Gates Foundation, 2013; Paufler & Amrein-Beardsley, 2013). As such, there appears to be little hope in reducing bias enough so that VAM-use could be realistically and practically relied upon for accurately capturing teacher quality.

Observations and Rubrics in Teacher Evaluations. Observation-based teacher evaluations have been a common practice in measuring teacher effectiveness for many years (Hill et al., 2012). The methods of observation-based evaluations became an issue of concern after the New Teacher Project's (2009) release of the "The Widget Effect."

The authors of the report asserted that school administrators were failing to differentiate between effective and ineffective teaching, stating:

> This report examines our pervasive and longstanding failure to recognize and respond to variations in the effectiveness of our teachers. At the heart of the matter are teacher evaluation systems, which in theory should serve as the primary mechanism for assessing such variations, but in practice tell us little about how one teacher differs from any other, except teachers whose performance is so egregiously poor as to warrant dismissal, (Weisberg et al., 2009, p. 4).

In the new policy era, teacher observations continue to hold a significant place in teacher evaluation systems. RttT required applicants to develop multi-measure systems that included classroom observations (RttT, 2011). Multi-measure systems have been recommended as the fairest and most effective method for capturing the complexities of teaching (Kane & Staiger, 2012). Observation instruments that allow the observer to collect evidence are thought to be the most appropriate way for conducting an observation (Darling-Hammond, 2013; Guskey, 2002; O'Malley et al., 2003; Simon & Boyer, 1969; Van-Tassel, Quek, & Feng, 2007), such as rubrics with predetermined objectives that are set to numerical values.

As per RttT, observations should be used for (1) fair evaluative purposes and (2) providing thorough feedback for instructional improvement. Until recently, teacher evaluations were not often used for professional development purposes (Brandt, Mathers, Oliva, Brown-Sims, & Hess, 2007; Ellett & Garland, 1987; Loup, Garland, Ellett, Rugutt, 1996). As such, there are recommendations that schools can follow to develop observation practices that can be used both for evaluative and professional development

purposes. The first consideration should be with the evaluators. While the common practice has been for administrators to conduct all teacher observations (Brandt et al., 2007), the new recommendation is that mentors and peers, as well as administrators, observe teachers (Oliva, Mathers, & Laine, 2009). This enables observers with similar experiences (e.g., content knowledge and instructional background) who teach similar students to provide feedback for teachers who might not get it otherwise (Goldstein & Noguera, 2006). Regardless of who conducts the observations, experts also recommend that the observers be well trained in order to increase inter-rater reliability and decrease rater bias (Darling-Hammond, 2013; Hill et al., 2012; Loup et al., 1996; Oliva et al., 2009; Stiggans & Duke, 1988).

Experts also recommend the modification of the frequency of observations (Oliva et al., 2009). Oftentimes, tenured teachers are only observed once every few years (Brandt et al., 2007; Sweeney & Manatt 1984). However, if observations are to be used for constructive feedback and, subsequently, improved instructional practices, then evaluators should observe tenured and untenured teachers more than once every school year (Blunk, 2007). Frequent observations might also help increase reliability; however, it cannot be guaranteed that any specific number of observations would satisfy the level of reliability needed to make consequential decisions (e.g., tenure, merit pay, termination; Hill et al., 2012). For this reason, multiple measures of teacher performance hold as the highest recommendation for teacher evaluation systems across the board.

TAP: The System for Teacher and Student Advancement. TAP is one of the nation's leading comprehensive teacher evaluation systems, and it is the system at play in the context of this study. While I will describe, in detail, the specific and relevant

elements of TAP in the "Policy Framework" section later in this chapter, I will use this space to discuss the research related to TAP. First, TAP was developed by Lowell Milken and the Milken Family Foundation and was first implemented in the 2000-2001 school year (Daley & Kim, 2012). NIET, the coordinators of TAP and a 501(c)(3) public charity, primarily works with high-need schools and have formed partnerships with four states (Louisiana, South Carolina, Indiana, and Texas), as well as Arizona State University. As of the 2011-2012 school year, 80 districts serving approximately 347 schools, 20,000 teachers, and 200,000 students, had adopted the TAP teacher evaluation system. TAP advertises its alignment with the TIF grant competition, stating that: "In the last round of TIF funding [2010], applicants proposing the TAP system won eight of the 34 awarded grants," (TAPsystem.org, 2010). The report goes on to describe the specific TAP components that align with the TIF expectations. The following sections will describe the TAP approach to meeting the state and federal demands for teacher evaluations in the Osborn School District in Arizona.

Nearly all of the research on TAP, or TAP-related sites, has been funded by internal entities. NIET, the sponsor of TAP, has funded numerous studies and reports. For example the TAP website includes 32 (plus an additional research summary report) articles or reports in support of TAP. Of those, NIET produced 10 of the reports; and TAP, prior to joining NIET, produced two. NIET provided the sole funding for an additional three reports; and the Milken Family Foundation, the founder of TAP, funded eight. The Joyce Foundation, one of TAP's leading financial supporters (tapsystem.org, 2013), funded four of the reports, while the Algiers Charter School Association, that receives TIF funding for their TAP system, produced two of the reports. Only two of the

reports were not directly written or funded by a TAP entity. Sally Hudson (2010) wrote one of the reports for her undergraduate honors thesis paper; and The Center for High Impact Philanthropy (2010) wrote the second one, which was a brief about ways philanthropists can get involved in improving teacher quality. The authors mentioned TAP once as an "organization working in the area [of] comprehensive teacher evaluations that are linked to ongoing professional development and distribution of teachers," (p. 3). The authors do not explain why or how TAP is an exemplary model of a teacher evaluation system, nor do they cite any other studies. Of the 29 total reports, none were peer-reviewed.

External research on TAP is scarce, as is the research regarding TIF sites in general. Schacter and Thum (2005) employed a multivariate multi-level model to explore the impact of a TAP evaluation system implementation on student achievement. They found that schools that used TAP showed significant growth in student achievement, but that growth varied by school and fidelity of implementation (i.e., schools that adhered strictly to the TAP system demonstrated greater growth than schools that did not). Mathematica Policy Research, Inc. released a report on the impact of TAP in Chicago Public Schools (Glazerman & Seifullah, 2012). The authors looked at Chicago TAP versus non-TAP schools over a four-year period and found that (1) teachers in TAP schools experienced more opportunities for mentoring; (2) there was no significant difference in school climate or teacher attitudes; (3) there was no significant increase in student achievement in the TAP schools; (4) there was some indication of increased teacher retention, but it was not universal across schools, cohorts, or subgroups.

In March 2014, NIET released a summary of TAP research. According to the

authors of the report, TAP has succeeded in the following areas: (1) can differentiate between levels of teacher effectiveness; (2) can provide feedback for improvement; (3) promotes and allows for data-driven professional development opportunities; (4) encourages and allows for recruitment and retention of effective teachers; (5) creates a collegial environment focused on student learning (Barnett, Rinthapol, & Hudgens, 2014).

The Outcomes Associated with Teacher Evaluation Policies

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

Collins (2012) sought the perspectives of the teachers via survey methods and found that teachers reported concerns with the reliability, validity, and bias of the VAM-use in their district. Additionally, the study suggested unintended consequences associated with the high-stakes use of the VAM, in which teachers admitted to teaching to

the test, targeting instruction to students most likely to show growth, and unwillingness to collaborate or share best practices with other teachers who were seen as competitors. While the unintended consequences were troublesome, equally as troublesome was that teachers also reported little to no use of VAM scores for making instructional decisions, thus raising the question whether the undergirding of VAM-based policies is to improve existing teacher quality or simply remove teachers from the profession. Assuming the former, teachers in the Collins study overwhelmingly stated that VAM reports were vague and unclear, and that they relied on other sources of data—not VAM data—to inform them of their teaching effectiveness.

While it might be too soon to expect more empirical work on the outcomes of these contemporary teacher evaluation policies, there have been legal cases that have resulted from questionable evaluation practices. For example, a group of Florida teachers filed a lawsuit in April, 2013 on the grounds of being evaluated based on students whom they do not teach (Jordan, 2013). Similar cases are likely to arise, as well as others due to the problems of reliability and validity with the VAMs that are currently used in state and district evaluation policies (Baker et al., 2013).

Conclusions on Literature

The literature suggests that the problem of a failing education system was first introduced during the Sputnik era of the 50s, reaffirmed in the 80s with the release of *A Nation at Risk*, and concretized in policy in the early 2000s with NCLB. RttT has joined its predecessors in addressing a now 60-year-old professed problem, this time directly targeting teachers as the root cause of failing schools. The main issue is that the targeted cause (e.g., poor teachers) of this problem has been supported with little (if any)

empirical evidence. Therefore, suggesting that another round of increased accountability mechanisms will do anything to improve the quality of the education system is increasingly showing to have negative consequential outcomes for teachers – while even less empirical evidence exists on how student achievement and learning outcomes have been impacted.

Rather, a majority of the literature suggests that even though teachers are the most significant in-school factor in student achievement scores (Goldhaber, 2002; Sanders, 2000), they really only account for approximately 10-20% of student achievement score variation overall (Kennedy, 2010; Gabriel & Allington, 2011; Xu, Ozek, & Corritore, 2012), and factors such as home-life, health, poverty, etc., things well beyond the control of teachers and schools, largely influence student achievement scores (Berliner, 2013). Thirty years of increased accountability policies have resulted in no evidence to suggest that more of the same will address the root causes of low student achievement scores (Au, 2009; Haney, 2000; Hursh, 2008; Klein et al., 2000; Orfield & Kornhaber, 2001).

Policy Framework

The purpose of the following sections is to define the teacher evaluation policy landscape within which Desert School District is located. Included is the tiered structure, starting with the federally initiated incentive packages (i.e., RttT and TIF programs), followed by Arizona's legislative framework for teacher evaluations as of September 2014, and then the local school district's adoption of the TAP system for teacher evaluations. Each tier will include the three major components as defined by RttT: observation procedures, the use of student assessment data to evaluate teachers, and personnel decisions based on the evaluations.

Tier One: Federal Incentive Programs. As of summer 2014, there have been no federal legislations on teacher evaluations, per se. However, there have been incentive programs that have de facto regulated the teacher evaluation systems across the United States. Similarly, the federal grant program, Teacher Incentive Fund (TIF), as well as the Elementary and Secondary Elementary Act (ESEA) Flexibility (i.e., NCLB waiver) have had an influence on contemporary teacher evaluation policies and practices.

Race to the Top. Under the American Recovery and Reinvestment Act (ARRA) of 2009, President Obama, along with a bi-partisan Congress, allocated $4.35 billion to education reform efforts. With the stated goal being to encourage innovation, the legislation itself did not mandate a prescribed set of guidelines for education reform. Instead, the initiative manifested in the form of a statewide competition, RttT, that called for proposals that met four major tenets: 1) adoption of the Common Core State Standards (CCSS), 2) a plan to develop a data system to inform parents, teachers, and schools about students, 3) a plan to recruit, develop, and retain effective teachers and principals, including an evaluation system to identify effective teachers and principals, and 4) a strategic plan to identify and turn around low-achieving schools (RttT, 2011).

By January 20, 2010, 40 states had submitted applications for the first phase of the competition. Applications were assessed on a point system with a maximum value of 500 points; and only two states, Tennessee (awarded $500 million) and Delaware (awarded $100 million), won grants in round one. Since then, 46 states and the District of Columbia have submitted applications for RttT funds; of those, 34 have revamped their education policies to meet the grant's four major principles (i.e., CCSS, improved data systems, effective teachers and principals, and turning around low-achieving schools). To

date, RttT has awarded more than four billion dollars to 21 states and the District of Columbia. Accordingly, states across the country have shaped new teacher evaluation systems to meet the demands of the RttT competition.

Following are the specific expectations of the state applications for the teacher evaluation system section of the competition application—a section worth 58 points (i.e., more than 10% of the overall application), which is the second most valuable section, trailing the overall education reform agenda section by only seven points. Within the teacher evaluation section, there are four subcategories: 1) measurement of student growth, 2) fair evaluations that differentiate teacher and principal effectiveness, 3) observations and feedback, and 4) informed decisions (e.g., personnel decisions) based on evaluations.

Measurement of Student Growth. RttT defined the measurement of student growth as to "establish clear approaches to measuring student growth (as defined in this notice) and measure it for each individual student," (section D(2)(i); p. 34, RttT Application, 2009). It defined student growth as: "the change in student achievement for an individual student between two or more points in time. A State may also include other measures that are rigorous and comparable across classrooms," (p. 11, RttT Application, 2009). This component of the teacher evaluation section was worth five of the 58 total points. States were required to provide percentages of participating local education agencies (LEA) that measure student growth at the time of the application submission, as well as the anticipated percentages for the four subsequent years. States were able to choose their method of measuring student growth (e.g., value-added models, student growth percentile models, etc.).

Fair Evaluation Systems that Differentiate Teacher Effectiveness. RttT defined the differentiation of teacher and principal effectiveness as: "Design and implement rigorous, transparent, and fair evaluation systems for teachers and principals that (a) differentiate effectiveness using multiple rating categories that take into account data on student growth as a significant factor, and (b) are designed and developed with teacher and principal involvement," (section D(2)(ii); p. 34, RttT Application, 2009). This component was worth 15 points. States had to submit the percentages of LEAs with qualifying evaluation systems for teachers, as well as principals, for the years of application and projected percentages for the four subsequent years.

Observations and Feedback. RttT defined observations and feedback as "Conduct annual evaluations of teachers and principals that include timely and constructive feedback; as part of such evaluations, provide teachers and principals with data on student growth for their students, classes, and schools," (section D(2)(iii); p. 34, RttT Application, 2009). This component was worth 10 points. There was no specific recommendation provided for evidence; however, it was listed that in the future, the state would have to provide the percentages of teachers who were identified as effective and ineffective in the previous academic year.

Use of Evaluation Outcomes to Inform Decisions. The use of evaluations to inform decisions was the most valuable component, totaling 28 of the possible 58 points. It was defined by (taken directly from RttT): 1) developing teachers and principals, including by providing relevant coaching, induction support, and/or professional development; 2) compensating, promoting, and retaining teachers and principals, including by providing opportunities for highly effective teachers and principals (both as

defined in this notice) to obtain additional compensation and be given additional responsibilities; 3) whether to grant tenure and/or full certification (where applicable) to teachers and principals using rigorous standards and streamlined, transparent, and fair procedures; 4) removing ineffective tenured and untenured teachers and principals after they have had ample opportunities to improve, and ensuring that such decisions are made using rigorous standards and streamlined, transparent, and fair procedures. (section D(2)(iv); p. 34, RttT Application, 2009).

Applicants had to provide the percentages of participating LEAs that use evaluation systems to make decisions based on: 1) professional development, 2) compensation, 3) promotion, 4) retention, 5) granting of tenure and/or full certification, 6) dismissal.

RttT recipients are also required to submit an annual report describing their progress on their reform efforts, including their teacher evaluation system. If the FDOE determines that the state is not up to standard, then the Department can take action, such as by withholding funds or requiring the state to reimburse rewarded funds. Additionally, the Institute of Education Sciences (IES) performs national surveys to evaluate the impact of the program (RttT Application, 2009).

Teacher Incentive Fund. Similar to the RttT competition, the Teacher Incentive Fund (TIF) grant competition was initiated to prompt school reform but with a specific focus on performance-based compensation systems (PBCS) for teachers and principals based on student growth and classroom evaluations in high-needs schools (TIF, 2010). The stated purpose of the competition was to increase teacher effectiveness and student achievement. The TIF program was originally authorized under the Departments of

Labor, Health and Human Services, and Education, and Related Agencies Appropriations Act, 2006, Title V, Part D. The competition was open for LEAs (including charters that were LEAs), states that partnered with one or more LEAs, or non-profits that partnered with one or more LEAs.

Since 2006, the federal government has award four rounds of TIF grants. For the purposes of this paper, the I will focus on the 2010 (i.e., cohort three) call for proposals, as this was the year for which Arizona State University, in collaboration with the district, which is the site of the proposed study, applied and received a $43.8 million TIF grant for the years 2010-2015. The funds for the 2010 competition were made from the Consolidated Appropriations Act of 2010 (Public Law 111-117) and the American Recovery and Reinvestment Act of 2009, Division A, Title VIII, Public Law No. 111-5 (TIF Application, 2010).

The 2010 TIF application included six priorities: 1) differentiated levels of compensation for teachers and principals based on effectiveness, 2) fiscal sustainability of a PBCS, 3) a comprehensive approach to the PBCS, 4) use of value-added measures of student achievement to evaluate teachers, 5) Increased Recruitment and Retention of Effective Teachers to Serve High-Need Students and in Hard-to-Staff Subjects and Specialty Areas in High-Need Schools, and 6) New Applicants to the Teacher Incentive Fund. Though all of the components impact teacher evaluation policies at the local level, three of the components directly specify the ways in which teachers should be evaluated: differentiated levels of compensation for teachers and principals based on effectiveness, a comprehensive approach to the PBCS, use of value-added measures of student achievement to evaluate teachers. These three priorities also align with the RttT

expectations.

Differentiation of Incentive Pay. TIF defined differentiation of teacher and principal effectiveness as: "an applicant must demonstrate, in its application, that it will develop and implement a PBCS that rewards, at differentiated levels, teachers…who demonstrate their effectiveness by improving student achievement as part of the coherent and integrated approach of the local educational agency (LEA) to strengthening the educator workforce," (p. 8, TIF Application, 2010). Specific requirements under this section are: a) evaluations must give significant weight to student growth data; and b) must include multiple classroom observations (minimum of two) conducted by a trained evaluator with a rubric; c) must include additional forms of evidence; d) ensure inter-rater reliability; and e) must show how the differentiated pay incentives were justified (e.g., aligned to differentiation of effectiveness).

Comprehensive Approach to PBCS. TIF defined comprehensive approach to PBCS as: "the applicant must provide, in its application, evidence that the proposed PBCS is aligned with a coherent and integrated strategy for strengthening the educator workforce, including in the use of data and evaluations for professional development and retention and tenure decisions in the LEA or LEAs participating in the project during and after the end of the TIF project period," (p. 9, TIF Application, 2010). Also, applicants had to demonstrate their plan to provide teachers with professional development on how to use the evaluation feedback to improve their instructional practices.

Value-Added Measures of Student Achievement. TIF defined value-added measures of student achievement as: "the applicant must demonstrate, in its application, that the proposed PBCS for teachers…will use a value-added measure of the impact on

student growth as a significant factor in calculating differentiated levels of compensation provided to teachers," (p. 9, TIF Application, 2010). The specific value-added model is at the discretion of the LEA, however, the LEA must demonstrate a plan to explain the model to teachers to enable them to make instructional decisions based off of the model's data.

Tier Two: State Framework for Arizona. In an effort to submit a competitive application to the Race to the Top competition, Arizona legislators proposed and passed Senate Bill 1040 (A.R.S. §15-203(A)(38)), mandating that the State Board of Education (SBE) develop a state framework for teacher and principal evaluations by December 15, 2011. Accordingly, SBE formed the Task Force on Teacher and Principal Evaluations, comprised of teachers (public and charter), principals, university professors, school board members, union representatives, and state affiliates, to develop the Arizona Framework for Measuring Teacher Effectiveness (2011). By the 2012-2013 school year, LEAs were required to stay within the framework, but had the flexibility to develop and implement their own evaluation systems. LEAs had to determine such things as the specific growth or value-added model to adopt, the rubric for which to use to conduct classroom evaluations on teacher performance, and the personnel decisions to be made based on the evaluation outcomes.

Arizona's first two attempts at the RttT competition were unsuccessful; but in 2011, Arizona applied for, and won, a phase three RttT grant for $25 million. Following the expectations of the grant, the Arizona Framework for Measuring Teacher and Principal Effectiveness provided LEAs with general guidelines for using student growth data and teacher performance indicators to evaluate teachers. Though not detailed in the

framework, further expectations for teacher evaluation systems were explicitly indicated in the Arizona RttT application. In the following sections, I will explicate the teacher evaluation system expectations as made clear by the Arizona Framework for Measuring Teacher and Principal Effectiveness, as well as the Arizona RttT application. I also reviewed Arizona's No Child Left Behind waiver application, yet I found no additional information regarding the state teacher evaluation framework (Arizona Department of Education, 2012).

Use of Student Growth Data. The Arizona Framework differentiates between those teachers who have students that take the Arizona Instrument to Measure Standards (AIMS), which are fourth through eighth grade reading/language arts and mathematics teachers (i.e., Group A), and those teachers who do not have students that take AIMS (i.e., Group B). Though the Framework explicitly states that LEAs are not required to use the AIMS as the measure for student growth calculations, they do recommend it as a valid and reliable assessment for Group A teachers.

Regarding evaluation calculations, Group A evaluations must include a 33% to 50% weight of classroom-level student growth data. The Framework does not mandate a particular value-added or growth model for LEAs to use in their teacher evaluation systems; however, as included in the RttT application, the Arizona Department of Education does calculate student growth using the Arizona Growth Model, which is an adaptation of the Colorado Growth Model developed by Bettebenner (2011). "Growth" (i.e., learning) is measured by placing students into similar testing peer groups, determining the expected growth of the group (i.e., one year's growth), and dividing the group into quintiles (i.e., 1=significantly less than one year's growth, 2=less than one

year's growth, 3=one year's growth, 4=more than one year's growth, 5=significantly more than one year's growth). The Arizona Growth Model does not include any covariates to account for outside factors (e.g., socioeconomic status, race, etc.). To determine the teacher's value-added score at the classroom level, after each student's individual growth has been determined, the students of the teacher are rank ordered, and the median growth score of the class is the teacher's value-added score. The same process is used to calculate the school's value-added score, which is how Group B teachers are evaluated. Since these teachers do not necessarily have what the state considered a valid and reliable measure of student achievement, 33% to 50% of Group B teachers' evaluations are comprised of school-level growth. School-level growth can be included in Group A teachers' evaluations, but can only account for up to 17% of the total evaluation.

Fair Evaluation Systems that Differentiate Teacher Effectiveness. The Framework does not specify the way in which LEAs should differentiate teacher effectiveness. The RttT application requires that LEAs develop evaluation systems that include four levels of effectiveness and provides the following examples: highly effective, effective, minimally effective, and ineffective. The application also included a mandate that LEAs must include teachers and principals in development and improvement of the evaluation instruments. Also, evaluators must attend professional development that certifies them to fairly evaluate teachers.

Observations and Feedback. The Framework requires that evaluators perform multiple classroom observations throughout the year, though a specific number of observations is not indicated. Evaluators are required to use rubrics that are based off of the national teaching standards. The "Teacher Performance" component of the total

teacher evaluation calculation must be between 50% and 67% for Group A and Group B teachers. The RttT application does not explicate anything for the observations and feedback besides granting authority to the SBE via the Framework.

Use of Evaluation Outcomes to Inform Decisions. The Framework does not indicate requirements for decisions to be made on evaluation outcomes. However, the RttT application included four strategies that the state would adopt to ensure LEAs were making informed decisions based on the evaluation outcome data (taken directly from the Arizona RttT application): 1) ensure that evaluation results are used to develop teachers and principals to increase their instructional effectiveness; 2) encourage use of evaluation results to compensate, promote, and retain effective teachers and principals; 3) ensure that evaluation results inform the granting of full certification to teachers and principals using rigorous standards and streamlined, transparent and fair procedures; 4) ensure that evaluation results are used to inform the removal of ineffective continuing and non-continuing teachers and principals after they have had ample opportunities to improve, and ensure that such decisions are made using rigorous standards and streamlined, transparent, and fair procedures, (p. 143-145, Arizona RttT Application, 2010).

Tier Three: Local Framework for Desert School District. In 2010, Desert School District joined a team of 12 high-needs, Arizona school districts along with Arizona State University to apply for a TIF grant. The project, called the Arizona Ready-for-Rigor Project, acquired a $43.8 million grant to be used for their proposed five-year plan to implement a performance-based compensation system (PBCS). The Arizona Ready-for-Rigor Project partnered with the National Institute for Excellence in Teaching (NIET) to implement the TAP system (formally the Teaching Advancement Program),

which is a comprehensive teacher evaluation system that focuses on four primary areas: 1) multiple career paths, 2) ongoing applied professional growth, 3) instructionally-focused accountability, and 4) performance-based compensation systems. Given Desert's participation in the Arizona Ready-for-Rigor Project, the following explication of the district's teacher evaluation system will be contextualized within the TAP framework.

TAP: System for Teacher and Student Advancement. TAP was developed by Lowell Milken and the Milken Family Foundation and was first implemented in the 2000-2001 school year (Daley & Kim, 2012). NIET, the coordinators of TAP and a 501(c)(3) public charity, primarily works with high-need schools and have formed partnerships with four states (Louisiana, South Carolina, Indiana, and Texas), as well as Arizona State University. As of the 2011-2012 school year, 80 districts serving approximately 347 schools, 20,000 teachers, and 200,000 students, had adopted the TAP teacher evaluation system. TAP advertises its alignment with the TIF grant competition, stating that: "In the last round of TIF funding [2010], applicants proposing the TAP system won eight of the 34 awarded grants," (TAPsystem.org, 2010). The report goes on to describe the specific TAP components that align with the TIF expectations. The following sections will describe the TAP approach to meeting the state and federal demands for teacher evaluations in the Desert School District in Arizona.

Use of Student Growth Data. TAP's evaluation calculation is similar to that of Arizona's framework in that it divides teachers into two groups based on the possibility to calculate classroom-level value-added scores. That is, teachers of students who take the annual AIMS assessment (i.e., 3rd-8th and 10th grade English/language arts and mathematics teachers), have a different evaluation calculation than teachers who do not.

Teachers with classroom-level value-added data (i.e., Group A) have a breakdown of 30% classroom-level value-added, 20% school-level value-added, and 50% skills, knowledge, and responsibilities (e.g., teacher observation scores); while teachers without classroom-level value-added data (i.e., Group B) have a breakdown of 50% school-level value-added and 50% skills, knowledge, and responsibilities.

Desert School District, in conjunction with the rest of the Ready-for-Rigor Project, chose to use the Arizona Growth Model (i.e., Colorado Growth Model, Bettebenner, 2011). The Ready-for-Rigor Project provided districts with some flexibility in determining the way they labeled their teachers. Given that Desert is a K-8 school district, a large percentage of their teachers fit in Group A. However, there are some teachers (e.g., K-2 teachers, specials teachers, English Language Learner teachers, middle school content other than English/language arts or mathematics teachers, etc.) who do not have students who take the AIMS and, thus, might fit into Group A or Group B. One student can be attributed to up to four teachers. Therefore, if a district so chooses, a seventh-grade student's growth on the English/language arts test might be used to calculate not only the English/language arts teacher's classroom-level value-added score, but also the social studies teacher's value-added score. A similar process can take place in other special circumstances, such as teachers who team-teach or switch students halfway through the year. Each district is capable of making the decision about grouping teachers into A or B. Desert has chosen to equally weight teachers' value-added scores (personal communication with Patricia Tate, Assistant Superintendent at Desert, July 30, 2013). For example, if a student has one primary teacher and one reading specialist, the student's growth score is included in both of the teachers' value-added calculation. The only

teachers who fit into Group B are those who do not teach any students who are growth score eligible (e.g., kindergarten-only teachers).

Fair Evaluation Systems that Differentiate Teacher Effectiveness. NIET recommends that schools gain a 75% approval rating from faculty before adopting the TAP system in order to build collegiality and active participation. The TAP system attempts to foster collegiality by encouraging collaboration through master, mentor, and career teachers. TAP refers to this as "multiple career paths," and it is one way that schools differentiate effectiveness of teachers and provide support for teachers to continually improve their practice. Based on the evaluation outcomes, teachers have the ability to move up the "ladder" to other positions that allow them to coach other teachers. Master teachers are released from their regular teaching duties to spend their on 1) leadership team participation, 2) research, 3) cluster group planning and implementation, 4) individual growth plan management, 5) evaluations and conferencing, and 6) classroom follow-ups (p. 11-12, TAP System Leadership Handbook). Mentor teachers have similar duties as the master teachers, but at a lesser amount. They spend more time teaching students in the traditional sense and less time on coaching responsibilities. Mentor and master teachers have a range of teaching experiences, as the identification of such positions are not based on seniority, but rather on merit according to the TAP evaluation system.

Teacher effectiveness is differentiated based on a five-point scale, ranging from unsatisfactory to exemplary. Each criterion (i.e., classroom-level value-added, school-level value-added, and skills, knowledge, and responsibilities) is based on this scale. The skills, knowledge, and responsibilities component is measured by a five-point rubric,

while the classroom and school-level value-added components are measured by the average growth of the students (i.e., 1=significantly less than one year of growth, 2=less than one year of growth, 3= one year of growth, 4=more than one year of growth, and 5= significantly more than one year of growth). The total rating of a teacher is calculated by the evaluation formula based on the teacher's affiliation with Group A (i.e., classroom-level VA = 30%, school-level VA = 20%, and SKR = 50%) or Group B (school-level VA = 50% and SKR = 50%).

*Observations and Feedback***.** The TAP system has a rigid process for observations and feedback, which falls under the SKR criterion. Before administrators or master and mentor teachers are permitted to evaluate teachers, they must complete a 35-hour training course and pass an online certification exam, which must be renewed annually. The primary focus of the training course is to familiarize potential evaluators with the TAP observation and conference protocols. During the training course, evaluators learn about the rubrics and the indicators, as well as how to collect evidence to justify evaluative decisions. The evaluators participate in various mock-observations by watching full-length, videotaped lessons, collecting evidence, and scoring the teachers' lessons based on the evidence. Evaluators also observe pre- and post-conferences, as well as practice conferences with other evaluators employing specific, TAP-recommended reflection questions.

Pre-conference. There is a standard practice for teacher observations and feedback under the TAP system. Prior to announced observations, teachers submit a lesson plan and participate in a pre-conference with the evaluator. The evaluator conducts the conference based on a specified TAP model that includes pre-determined reflection

questions. The goal of the evaluator is to guide the teacher through a series of self-reflection questions that allows the teacher to think about his/her instructional decisions. Examples of pre-conference reflection questions are:

1. What are the pre-requisite skills needed for students to be successful?
2. What changes or adjustments will you need to make if students do not show evidence of mastery of the sub-objectives?
3. How will you know students have mastered the objectives?
4. Is there anything you want me to be aware of before the observation?
5. How will you differentiate your instruction to address various learning styles?

Observation. It is recommended that evaluators start with a general question and narrow to more specific questions as the conference proceeds, fostering a conversation of metacognitive reflection on the teacher's part. Following the conference, the evaluator is required to observe an entire lesson, regardless of time length. During the lesson, the evaluator should take copious, objective notes on the teacher's talk, behavior, materials, and practices, as well as the students' talk and behaviors. Evaluators are encouraged to capture as much of the lesson as possible. When the lesson is finished, the evaluator should spend time (recommended one hour) sifting through the evidence and using the rubric to evaluate the teacher's performance and planning for the post-conference.

The teacher performance rubric is comprised of four separate components: instruction rubric, learning environment rubric, designing and planning instruction rubric, and professional responsibilities. The instruction rubric is comprised of 12 categories: standards and objectives, motivating students, presenting instructional content, lesson structure and pacing, activities and materials, questioning, academic feedback, grouping

students, teacher content knowledge, teacher knowledge of students, thinking, and problem solving. Each category has a number of indicators that teachers must meet in order to earn proficient or exemplary standings. For example, the objectives and standards category contains six indicators (i.e., most learning objectives and state content standards are communicated, sub-objectives are mostly aligned to the lesson's major objective, learning objectives are connected to what students have previously learned, expectations for student performance are clear, state standards are displayed, there is evidence that most students demonstrate mastery of the objective) that teachers must meet to be marked as proficient in that category. The levels (e.g., unsatisfactory, proficient, exemplary) are differentiated with qualifying terms (e.g., few, most, all, etc.). With the exception of the thinking and problem solving categories, there are a total of 53 indicators that teachers must meet during the single lesson in order to be marked as proficient on the instruction rubric. In regards to the thinking and problem solving categories, evaluators should collect evidence from multiple observations before making conclusive decisions about the proficiency level of the teacher.

The Designing Instruction rubric is comprised of three separate categories (i.e., instructional plans, student work, and assessment) for a total of 15 indicators in the proficiency column. For example, a proficient teacher in instructional plans would demonstrate: 1) Goals aligned to state content standards; 2) Activities, materials, and assessments that: a) are aligned to state standards, b) are sequenced from basic to complex, c) build on prior student knowledge, d) provide appropriate time for student work, and lesson and unit closure; 3) Evidence that plan is appropriate for the age, knowledge, and interests of most learners and; 4) Evidence that the plan provides some

opportunities to accommodate individual student needs.

The Learning Environment rubric is comprised of four categories (i.e., expectations, managing student behavior, environment, and respectful culture) for a total of 17 indicators in the proficient column. The Responsibilities component of the performance evaluation is not a rubric, but a set of responsibilities for which teachers are expected to hold. These responsibilities vary depending on the teacher's position as a career, mentor, or master teacher.

Post-Conference. After the evaluator has evaluated the lesson based on the evidence collected during the observation, a post-conference is held with the evaluator and the teacher. Before the conference, the evaluator scripts a plan for the meeting, including reflection questions for the teacher regarding an area of reinforcement (i.e., a practice that the teacher performed well and should continue) and an area of refinement (i.e., a practice that a teacher should work on in the future). The evaluator should also include time for an overall reflection of the lesson, as well as a time to review the scores from the rubric.

According to TAP, school districts are to decide the number of announced and unannounced[7] observations that should be made for each teacher. Desert School District has chosen to observe all career and mentor teacher four times a year by various evaluators (personal communication with Patricia Tate, Assistant Superintendent at Desert, July 30, 2013).

*Use of Evaluation Outcomes to Inform Decisions***.** The TAP system does not specify all of the decisions that should be made based on evaluation outcomes. However, it does emphasize the need to collect evidence in order to make informed decisions.

Desert uses evaluation outcomes for determining career, mentor, and master teachers, merit-based pay, and termination decisions. Along with researchers from Arizona State University, they are currently working to implement an appropriate assessment for their kindergarten through second grade students in order to include nearly all of their teachers in the teacher-level value-added eligible pool.

Conclusion

Teacher evaluation policies have taken root across the country, affecting the almost three million teachers in America's public schools, sometimes in highly consequential ways, despite the mounting research that says the accountability mechanisms are invalid, unreliable, biased, and unfair. Further, up to 70% of teachers nationwide cannot even be measured by the same instruments to which their counterparts are subjected. This problem is amplified by the fact that such teachers—the 70%—are subjected to evaluations that are determined based on students and content areas that they do not even teach (i.e., their VAM estimates are based on school-wide VAM estimates as based on students who do take the state standardized test).

Again, this is nothing new, as the U.S. has spent the past 30 years refining a series of accountability policies claiming to target the root cause of low educational quality. This has resulted in more than 30 years of failed policy and billions of federal dollars spent, leaving little to be expected from the next attempt. Such policies, by their very nature, have limited our scope of understanding the big picture problem masked as low educational quality. Policymakers have narrowed in so acutely on teachers, despite the limited impact that teachers ultimately have on student achievement scores (Kennedy, 2010; Gabriel & Allington, 2011; Xu, Ozek, & Corritore, 2012), so as to blindly ignore

that which has been shown to have the most profound impact on student achievement—poverty (e.g., Anyon, 2005; Berliner, 2006; Biddle, 2001).

CHAPTER 3

Theoretical and Methodological Framework

In this chapter, I provide the theoretical and methodological framework within which I developed and conducted the study. I start with a discussion of my theoretical transition from a critical structural approach to a poststructural approach. Then I detail how I have defined and operationalized discourse for the purpose of this study, followed by an explanation of Foucault's (1991) governmentality framework and related concepts. In the second half of the chapter, I link this framework to the shaping of a methodogical approach I used to answer the research questions.

Poststructuralism

At the onset of this study, I planned to use a critical ideological approach. Specifically, I was interested in using Critical Discourse Analysis (Fairclough, 1989; Fairclough & Wodak, 1997; Rogers, Malancharuvil-Berkes, Mosley, Hui, & O'Garro, 2005; van Dijk, 1993) to investigate how evaluative instruments mediated the evaluation process and thus legitimated evaluative decisions at DMS. Under this framework, I was held to the tenets of a structural ideological approach underscored by the assumption that society and reality could be understood and examined as a structure based on power relationships. Structuralists, for the most part, work from the epistemological belief that, given the appropriate analytical tools, the analyst can understand the system (or structures) from an external vantage point, thus granting them the ability to know how the system functions from a privileged perspective. For example, Marxism, feminism, and critical race theories are common critical structuralist approaches that seek to understand power relations based on concepts of class, gender, and race respectively.

However, soon after beginning the study, I found that such an approach left me with questions unanswerable by the analytical tools and theoretical assumptions with which I had chosen to work. I realized that the framework forced me to work from the assumptions that 1) power is a tangible thing that is held and used by a finite and determinable hegemonic group (e.g., the financial and political elite); 2) power relationships exist on a binary—the dominant versus the oppressed; 3) the participant responses reflect a "true" reality that can be investigated to reveal intentions, power, etc.; and 3) I, as the analyst, could remove myself from the system in order to make objective claims about who was telling the "truth." After the very first stages of data collection, I realized that these assumptions were too restricting and forced me, in a sense, to carry pre-conceived, deterministic ideas about how evaluation processes were operating at DMS. I realized, too, that power was operating in a much more complex way than what a CDA approach allowed me to understand. As such, I shifted my focus away from why-type questions and started to focus on how- and what-type questions—e.g., what is going on here? What conditions have to exist for these processes and practices to be made possible? How is discourse functioning here? These new questions led me away from a structuralist approach and towards a poststructuralist approach.

Poststructuralism—comprised of a heterogeneous group of theorists and methodologists (Peters, 1996)—manifested as a result of a philosophical shift in the late 1960s when a group of theorists began to refute the idea that society was made up of a set structures that hinged on hegemonic power relations (Marshall, 2004). Poststructuralists, though still interested in concepts like power, tend to focus less on the binary nature of an oppressed-dominant relationship, and seek rather to understand how discourses work to

shape reality and the knowable. Some of the renowned poststructuralist thinkers have been Jacques Derrida, known for his work in deconstruction (see Derrida, 1976; 1981; 1984); Jean-Francois Lyotard, known for his work on the "postmodern condition" (Lyotard, 1999); and Michel Foucault, known for his work in power/knowledge and governmentalities (Foucault, 1977; 1980; 1984; 1985; 1991). For the purposes of this study, I will be using a Foucualdian (1977; 1979) framework of governmentality (Foucault, 1991), while calling on various scholars who have also worked within this vein (Bacchi, 2000; Ball, 1990; 1993; 2003; Dean, 1994; Hacking, 1999; McWilliam & Jones, 2005; Rabinow, 1991; Rabinow & Rose, 2003; Rose, 1991, 1996, 1999). In the following sections I will develop my theoretical and methodological framework and define the concepts and assumptions upon which my study was founded and conducted.

Discourse

"Discourse" has been defined and operationalized in many different ways across disciplines, philosophies, and methodological approaches (Bacchi, 2000). For example, theorists who take on a structuralist perspective, such as critical discourse analysts, seek to understand how power is shaped by discourse, and how discourse reproduces power (Fairclough, 1989; Fairclough & Wodak, 1997; Rogers et al., 2005; van Dijk, 1993). In this vein, power can manifest in beliefs, policies, norms, behaviors, etc., which Gramsci referred to as "hegemony" (1971). Racism, classism, and sexism are all common forms of discrimination that are resultant of power dynamics. Here, also, language is assumed to be a structured symbolic representation of reality. Thus the role of the discourse analyst is to understand how the language (and sometimes the non-verbal cues and actions) fit into a larger social context, or narrative, allowing them to locate power and the process by

which power is reproduced. Poststructuralists, on the other hand, move away from the assumption that language represents reality and towards an assumption that language shapes reality by shaping possibilities and the knowable—at which point, power is problematized in terms of what is able to be said, by whom, and with what authority (Rose, 1999).

While texts often serve as the unit of analysis for CDA theorists, texts serve a different analytical purpose in poststructural discourse analyses. According to Foucault (1970):

> The outlines of a book are never clearly and stringently defined: no book can exist by its own powers; it always exists due to its conditioning and conditional relations to other books; it is a point in a network; it carries a system references—explicitly or not—to other books, other texts, or other sentences; and the structure of reference, and thereby the entire system of anatomy and heteronomy, depends on whether we are dealing with a dissertation on physics, a collection of political speeches, or a science fiction novel. It is true that the book presents itself as a tangible object; it clings to the tiny parallelepiped surrounding it: but its unity is variable and relative, does not let itself be constructed or stated and therefore cannot be described outside a discursive field (p. 152, as cited in Andersen, 2003).

Relatedly, language is historically and socially constructed, and thus language does not name things that exist in reality, but rather, the act of naming actually makes certain things possible, knowable, doable, etc. A relevant example might be that the teacher who does or does not add value to student learning is a specific type of teacher—that is, this type of teacher is defined in these terms (e.g., the teacher is labeled effective if s/he "adds

value" as measured by student achievement scores), which has real implications for what it means to be a "quality teacher." The "quality teacher," defined in this example, was unknowable or unimaginable before quality was (re)constituted by the idea to measure teaching based on student achievement scores. While I will discuss this particular example in depth in Chapters 4 and 5, the most important thing to point out here is that I am working from the assumption that discourses make and define possibilities, and through this process, certain ways of thinking and doing are made available. More simply put, discourse can be thought of as the knowable and the imaginable, and in order to make sense of discourses, one must seek to historicize and/or deconstruct the ways in which the "knowable" has come to be. Further, certain ideas are constituted as truth, or what Foucault (1980) referred to as "regimes of truth" (p. 131) rather than truth itself (McWilliams & Jones, 2005). In other words, what we think of as true (e.g., measures of teacher quality), though discursively constructed, is often accepted as truth as based on the rationality at play. For the purpose of this study, I will focus on the market-based discourse of the present era as it relates to teachers and teacher evaluation techniques. Thus "discourse" here relates to all the ways in which we have come to know about teacher, teacher quality, and the like as based on language, policies, practices, and instruments.

Accordingly, the role of the discourse analyst, from a poststructuralist perspective, is to understand how language (i.e., written and spoken), over time, has worked to shape some reality and constitute particular ways of knowing, doing, being, etc. Foremost, the researcher assumes that "no one stands [or can stand] outside discourse" (Bacchi, 2000, p. 45). Foucault, specifically, was interested in how discourses worked to produce

particular types of people as docile subjects (Foucault, 1984). He did not use conflict between the dominant and oppressed groups as his focal point, but rather he "[took] a series of oppositions—dividing practices involving men over women, of parents over children, of medicine over the population at large, of psychiatry over the mentally ill—as a starting point and attempt[ed] to define precisely what they have in common" (Peters, 1996, p. 82). As such, he sought to understand the relationship between power and the subject, or individual, by focusing on the conditions that create certain problems and solutions. Policy analysis, then, becomes less about trying to evaluate whether the policy addresses some problem, and more about how policies create, or give shape to problems in "the very proposals that are offered as responses" (Bacchi, 2000, p. 48).

Policy as Discourse. To think about policy, I will be calling specifically on Carol Bacchi's (2000) theorization of policy-as-discourse that argues that "the emphasis in policy-as-discourse analyses is upon the ways in which language, and more broadly discourse, sets limits upon what can be said" (p. 48), thought, and done. In this sense, policy works to define and constitute both solutions and problems. Put differently, policy does not solve some problem that already exists in reality. Rather, policy works to constitute problems and solutions because by specifying a solution, the *what* that the solution is trying to solve is "problematized" (Rabinow & Rose, 2003) and defined. A relevant example would be VAM-based teacher evaluation policies. The VAMs are meant to solve the problem of low teacher quality (either to improve low quality, or to identify and punish teachers who are of low quality), though in doing so, teacher quality is problematized only in terms of student achievement scores on standardized tests. This not only defines teachers as a problem, but it also confines the "problem" of teacher

quality in terms of the identified solution—student scores. I will expand on this idea in Chapter 4 when I discuss more specifically how teacher evaluation policies and practices problematize teachers and teacher quality.

This way of thinking about policy-as-discourse breaks from traditional policy approaches in that policy is not thought of as something that policymakers do; "policy-as-discourse approaches, by contrast, encourage deeper reflection on the contours of a particular policy discussion, the shape assigned a particular 'problem' (Bacchi, 2000, p. 48). As such, I will use this approach to analyze policy discussions about teacher evaluation policies in hopes of better understanding how teachers and teacher quality have been shaped by the very policies that been developed to evaluated them (discussed in detail in the Methods section of this chapter).

Discourse and the Subject. Within this framework, subjects (or individuals) are not discourse users, but rather are constituted by discourses (Bacchi, 2000; Burr, 1995). Accordingly, the question to ask is: how are subjects constituted, or defined, by discourses? Or, in the case of this study: how are teachers (and teacher quality) defined and constituted by evaluation policies, practices, and instruments? Foucault (2000) was interested in the relationship between discourses and the formation of subjects as determined (and re-determined over time) by types of knowledge available. He wrote that "what we should do is show the historical construction of a subject through a discourse understood as consisting of a set of strategies which are part of social practices" (Foucault, 2000, p. 4, as cited in Davies & Bansel, 2010, p. 5-6). Important to note here is that the subject is not a passive individual who has discourses done to them. Subjects are part of the discourse and part of the construction of themselves as subjects in relation to

the discourse (Dean, 1994).

Discourse and Power. Power was central to Foucault's work, yet he theorized power differently than that of structuralist philosophers who sought to locate power and power relations via discourses. First, Foucault and other policy-as-discourse theorists view power not as a tangible thing that someone (or group of someones) can possess and use against an oppressed group. Rather, power can be thought of in an omnipresent sense that is directly linked to power/knowledge production and infused in discourse (Rabinow, 1991). In fact, when referring to power, Foucault often used the term power/knowledge and related this to the idea of truth. Though, instead of thinking of truth as truth itself, he referred to this as "regimes of truth" that were constantly changing and being (re)negotiated over time (Foucault, 1980), writing:

> Each society has its regime of truth, its 'general politics' of truth: that is, the types of discourse which it accepts and makes function as true; the mechanisms and instances which enable one to distinguish true and false statements, the means by which each is sanctioned; the techniques and procedures accorded value in the acquisition of truth; the status of those who are charged with saying what counts as true' (p. 131).

Refuting the idea that society was made up of structures where power could be used as an instrument to keep oppressed groups down, Foucault saw power as being that which enables some things to be knowable and restricts others from not. "Foucault enabled us to see different kinds of relations between truth and power, in which power was a matter of the production of truth, and truth was itself a thing of this world, intrinsically bound to apparatuses like the prison, the hospital, the school and the clinic for its production and

circulation," (Rabinow & Rose, 2003, p. 3). The production of truth is also tied up in the way in which populations are governed, or the strategies by which populations can be turned into objects of knowledge and acted upon. These strategies of governance, or "governmentality" (Foucault, 1991), create the conditions upon which populations can be managed.

Governmentality

To define "governmentality," perhaps the best place to start is by defining that which it is not. It is not a study of governments in the way we might traditionally think of them (e.g., bureaucracy, official governing bodies, etc.). Instead it can be thought of as a combination of two words—"govern" and "mentality." "Govern" here refers to the way in which populations are controlled and produced, or, in other words to "structure the possible field of action of others' (Foucault, 1994, p. 341). Rose, O'Malley, and Valverde (2006) argued that "governmentality is far from a theory of power, authority, or even of governance. Rather, it asks particular questions of the phenomena that it seeks to understand, questions amenable to precise answers through empirical inquiry," (p. 85). "Mentality" here refers to the rationalities, strategies, and/or techniques that produce governable and self-governed persons. Put simply, the mentality can be thought of as the rationality upon which people can be controlled to behave in desired ways. Similarly, self-governance is the process by which subjects control themselves through various techniques (e.g., self-discipline, self-reflection, etc.) of good, civil, ethical behavior (Dean, 1999). As mentioned before, the "subject" is not a passive individual, but one who actively participates in an obedient society. Thus, the questioning of subjects' "conduct of conduct" is a key operating feature of modern societies (Foucault, 1982). To utilize a

governmentality approach in an analysis, Rose et al. (2006) argued that:

> Instead of seeing any single body—such as the state—as responsible for managing the conduct of citizens, this perspective recognizes that a whole variety of authorities govern in different sites, in relation to different objectives. Hence, a…set of questions emerges:
>
> - Who governs what?
> - According to what logics?
> - With what techniques?
> - Toward what ends? (p. 85).

For these reasons, to understand governmentality as the "conduct of conduct," one must first seek to understand the governing strategies, or rationalities, that make such conditions possible (Dean, 1994). I must point out, though, that "rationality" is not the same thing as legitimization in the sense that it acts to reaffirm an action already taken place. Rationality is, instead, built on some foundation of "truth" so as to "establish a kind of ethical basis for its actions," (Rose, 1999, p. 27). The question then becomes, who has the authority to make true statements, and how are these statements constructed?

Neoliberalism as a Governing Strategy. The mentalities (or rationalities) of governance have changed over time, but have always been present, starting from Ancient Greece to the contemporary neoliberal rationality (Lemke, 2002). Due to the time period of this study, the governing strategy that was of most relevance was neoliberalism. While scholars disagree on when neoliberalism came to be the primary form of governance, few disagree that it has been the governing strategy since no later than the mid 1980s (Peters, 1996; Rose, 1999). For the purposes of this study, I called upon Davies and Bansel's

(2010) use of neoliberalism to describe the academic setting:

1. All products are redefined in terms of their dollar values and their exchange value.
2. Through pitting individuals against each other in intensified competitive systems of funding with clearly defined measures of success, those individuals are de-individualized and converted into generic members of an auditable group (i.e., members are redefined in terms of quantifiable indicators and then held accountable to particular standards) (p. 6).

In other words, a neoliberal governing strategy is built on the notion that everything in a society (e.g., people, services, practices, etc.) can and should be quantified in terms of market value. Society functions on consumerism and responsibilized, ethical citizens. "In this new field, the citizen is to become a consumer, and his or her activity is to be understood in terms of the activation of rights of the consumer in the marketplace" (Rose, 1999, pp. 164–165). The onus is put on individuals to make responsible decisions about oneself and one's contribution to society. As such, key to neoliberalism is that of choice—or the subjects' freedom to choose responsibly. Inherent to choice is competition and thus a necessity for evaluating the worth of an object in terms of its market value. This process is done via various governing techniques, or technologies of governance.

Technologies of Governance. Governmentality is made up of two dimensions—rationality of governance (e.g., neoliberalism) and technologies of governance (Rose, 1999). Technologies of governance work on people to get them to behave in desired ways, which, in turn, produces desirable subjects, such as good workers, citizens, consumers, or in this case, teachers (Davies & Bansel, 2010).

> A technology of government, then, is an assemblage of forms of practical knowledge, with modes of perception, practices of calculation, vocabularies, types of authorities, forms of judgment, architectural forms, human capacities, non-human objects and devices, inscription techniques and so forth, traversed and transected by aspirations to achieve certain outcomes in terms of the conduct of the governed, (Rose, 1999, p. 52).

Put simply, technologies are the modes by which people are made subjects and objects of knowledge. By use of technologies, subjects are turned into objects of knowledge that can be acted upon. Then the technologies are used to control the conduct of subjects and to produce the desired types of subjects (e.g., responsible, ethical, civil, healthy, etc.).

Method

Local Context and Access

In 2010, Desert School District joined a team of 12 high-needs Arizona school districts along with Arizona State University to apply for a federal Teacher Incentive Fund (TIF) grant. The project, called the Arizona Ready-for-Rigor Project, acquired a $43.8 million grant to implement their proposed performance-based compensation system (PBCS). The Arizona Ready-for-Rigor Project partnered with the National Institute for Excellence in Teaching (NIET) to implement the TAP System for Student and Teacher Advancement (referred to as TAP from here on), which is a comprehensive teacher evaluation system that focuses on four primary areas: 1) multiple career paths, 2) ongoing applied professional growth, 3) instructionally-focused accountability, and 4) performance-based compensation systems.

Desert Middle School (DMS), specifically, is located in a metropolitan area of

Arizona and serves approximately 550 7th and 8th grade students. Many of their students do not speak English as their first language (19%), and almost all of their students qualify for free and reduced lunch (93%). DMS has defied most odds, however, as their school boasted a 74% passing rate on the state standardized test (AIMS) in reading and 65% in mathematics for the year of this study (2013-2014). These numbers are up from 69% and 59%, respectively, from the 2012-2013 school year. DMS also earned a B grade on Arizona's school grading system for 2013-2014. DMS is also unique in its teacher retention rate. While the turnover was greater right after TAP implementation (i.e., approximately 50%), the numbers of teachers leaving each year has dwindled to around 10% after the 2013-2014 academic year (all of these data are from personal contact with the Superintendent).

As for my choice in DMS for this study, I had worked for the TIF grant as a research assistant, where I was able to build a relationship with the then Assistant Superintendent (now Superintendent) and Curriculum Specialist (now Assistant Superintendent) of Desert School District. In exchange for some consultation work, they agreed to allow me into their middle school to conduct my study. This district was also an appropriate choice given its experience with the evaluation system. The year I collected data was the district's fourth year using TAP. As such, all of their teachers had been fully trained on TAP protocols; also, all of the administrators and evaluators (i.e., those who I interviewed) had been at Desert Middle School since TAP's inception. After receiving permission from the district, I applied for and received approval from Arizona State University's Internal Review Board to begin the study. At this time, I began to collect the data necessary for answering my research questions via a two-way analytic approach, as

described below.

A Complementary Approach

> Governmentality is far from a theory of power, authority, or even of governance. Rather, it asks particular questions of the phenomena that it seeks to understand, questions amenable to precise answers through empirical inquiry, (Rose et al., 2006, p. 85).

For this study, I sought to understand how evaluation policies and practices work to define teachers and teacher quality. I also sought to understand how the teachers have taken up this discourse in order to think about themselves and qualify their practice, quality, and worth accordingly. To this end, I approached the study using two complementary techniques. First, I collected and analyzed the policies, practices, and instruments associated with the teacher evaluation system at Desert Middle School in order to demonstrate the local manifestation of practices and instruments that have resulted from a neoliberal governing strategy. Here, I was interested in how such methods work to constitute teachers and give shape to the construct, teacher quality. Second, I interviewed teachers and their evaluators at the school. This second approach allowed me to get at how teachers have taken up and embodied the discourse as a means to think about themselves as teachers. I was interested, specifically, in how they have come to define themselves and their teaching quality in terms of their market value.

More specifically, I started with the idea that teacher evaluation policies have been argued on the assumption that teachers are in need of being observed, evaluated, and disciplined, and thus constituted as being inept to perform well without such management. Therefore, I started with the question of: how has teachers' conduct been

called in to question, or problematized, in the first place (Dean, 1999, p. 27)? Or, how have teachers been positioned as a problem in need of being fixed? To answer this question, I collected both official and unofficial policy documents (specified below) to trace the positioning of teachers as the problem in public schools, while also questioning *how* teachers have been problematized (i.e., what about teachers and/or teacher quality has been defined as the problem, and *how* has this problem been defined by the purported solutions?). While collecting documents, I kept detailed notes, noticing how teachers, and particularly teachers' conduct, was talked about in the pieces. I also analyzed the documents, as well as the field notes that I took during the evaluator training course, to get an understanding of the practices and instruments utilized to govern and discipline the teachers at Desert Middle School. Here I was particularly interested in how such procedures work to define teachers and teacher quality, and thus produce particular types of teachers. Then, in an effort to understand how teachers have taken up a neoliberal discourse, I talked to them as well.

For the teacher-related piece of the study, I used interview data to link the policies, practices, and instruments to the way in which the teachers and their evaluators in one middle school setting have come to see themselves as subjects in relation to such evaluations methods. I interviewed teachers and their evaluators (i.e., peer evaluators and school-based administrators) about their experiences with a multi-measure teacher evaluation system. In the next sections, I will discuss specific data collection and analyses procedures.

Part I: Analysis of Policies, Instruments, and Practices

Data and Data Collection. I collected official and unofficial documents related to

teacher evaluation policies and practices. This included, but was not limited to: policy statutes, promotional materials, political speeches, official position statements, and all other relevant and available literature on the official US Department of Education website (i.e., ed.gov) and the TAP system website (i.e., TAPsystem.org). For a list of all documents included in the analysis, please refer to Appendix A. In an effort to break away from a more traditional critical perspective that is focused on revealing an ideological "cause," and to move towards a poststructural framework that is more focused on how particular beliefs and practices might demonstrate various techniques of governance, I also collected data on the manifestation of such policies as they have appeared in the local context at Desert Middle School. To do so, I collected data on the practices and instruments utilized at DMS (e.g., rubrics, SGP protocol information, conference forms, etc.). I also attended the 35-hour TAP evaluator certification course, where I took field notes and collected evaluator training materials. Appendix A also includes all data for this part of the analysis.

Data Analysis. For this part of the analysis, I was interested in the notion that practices "systematically form the objects of which they speak; they do not identify objects, they constitute them and in the practice of doing so conceal their own invention' (Foucault, 1977, cited in Ball, 1990, p. 17). I first collected and read through each of the documents. During the first round of coding, I utilized holistic coding (Dey, 1993) to get an idea of the scope of the data. I also used this process to determine which documents were relevant and which were not. During the second cycle of coding, I paid more attention to the policies, practices, and instruments as means of technologies of governance and ways of problematizing both teachers and teaching quality. Throughout

the data collection, coding, and analysis stages, I strictly adhered to Saldaña's (2013) advice on analytic memo writing: "whenever anything related to and significant about the coding or analysis of the data comes to mind, stop whatever you're doing and write a memo about it immediately," (p. 33). This was my way of tracking my ongoing sense-making and theorizing (see Appendix B for a sample of my analytic memos). Most importantly, the memos served as a way for me to narrow my thinking from high-level observations and questions, to patterns and trends, to, finally, specific inferences about the data.

For the first round of data analysis, I paid specific attention to how teachers were positioned as "risky" in policy discussions and documents (Foucault, 1985; McWilliams & Jones, 2005). To do this, I focused on how teachers were described, as well as how they were presented as solutions to particular problems, or how they were presented as problems themselves. For the second round, I focused on how system procedures and tools were either suggested or legislated to help manage the potential risk that teachers presented. In other words, I was interested in the solutions to the problems of teachers and teacher quality. For the second round of analysis, I paid particular attention to the way in which practices and instruments were positioned as a means for managing, disciplining, and controlling teachers' conduct. For this part, I utilized a governmentality lens with a focus on the technologies of governance (Foucault, 1977; Rose, 1999).

To do this, I created a list of all mechanisms, techniques, practices, and the like that are used in the evaluation process (e.g., value-added models, rubrics, conferences, observations). Then I determined how each of these techniques functioned as mechanisms of governance (i.e., behavior control or discipline). Finally, I mapped the

techniques onto Foucault's (1977) and Rose's (1999) technologies of governance in order to draw conclusions about how the evaluation practices are working to govern teachers' behavior and define teacher quality.

Part II: Analysis of Teacher Interviews

Data and Data Collection. In order to get a better understanding of how teachers and their evaluators at one school have taken up a neoliberal discourse in terms of how they define themselves, their teaching quality, and their worth, I conducted in-depth, qualitative interviews (Spradley, 1979) with 11 participants at DMS. The participants included classroom teachers (N=7), peer evaluators (i.e., master teachers) (N=2), and school-based principals (N=2). The participants covered a wide range of content areas, grade levels, years of experience, and professional backgrounds (see Table 1). In one area, the participants lacked diversity, and that was of race. All of the participants were Caucasian, which is of particular importance given the demographics of Desert Middle School that predominantly serves Latino/a students. I prefer to have had a better representation of different races; however, a predominance of the teachers at Desert Middle School are Caucasian, and no other teachers volunteered to participate in the study. In future studies, it will be important to include participants of different races.

Table 1

Participant Characteristics

Pseudonym	Gender	Position	Grade Level	Content Area	Years in Position	TFA Y/N
Christina	F	Career Teacher	7th & 8th	Band	1	N
John	M	Career Teacher	8th	Literacy & Social Studies	2	Y

Mary	F	Career Teacher	7th & 8th	Art	5	N
Jennifer	F	Career Teacher	8th	Literacy	20	N
Sarah	F	Career Teacher	7th	Science	1	Y
Nicole	F	Career Teacher	8th	Mathematics	6	N
Melissa	F	Career Teacher	7th	Mathematics Special Education	1	Y
Robert	M	Master Teacher	8th	Literacy Honors	3	N
Heather	F	Master Teacher	8th	Mathematics	3	Y
Lisa	F	Vice Principal	NA	NA	4	N
Becky	F	Principal	NA	NA	10	N

To recruit participants, I briefly presented my proposal and request to the teachers during a morning staff meeting. When I first started the study, my intention was to only focus on teachers who received teacher-level value-added scores (Group A teachers). However, after receiving interest from several Group B teachers, I readdressed my research questions and contemplated my study's purpose, and I concluded that the inclusion of both Group A and Group B teachers would be not only acceptable, but would add a needed depth to the study that I might have missed otherwise. In all, seven (of 33) career teachers, two (of two) master teachers, and two (of two) administrators agreed to participate. I interviewed each of the teachers twice—once during their second cycle of

evaluations (during the fall semester) and again during their third or fourth cycle of evaluations (during the spring semester). I interviewed each of the evaluators once because of their limited availability. Each interview lasted approximately 45 minutes. The interviews were semi-structured and open-ended (Kvale, 1996; Spradley, 1979), with a focus on teachers' roles, responsibilities, and experiences as subjects of the evaluation system. In an attempt to build trust and openness, I structured the interviews as a conversation, while allowing the participants to co-construct the interview (Kvale, 1996). At the root of each interview, I had a set of core questions that I purposely asked all participants. Following Seidman's (2013), recommendation I structured the two interviews so that the first would focus on concrete experiences with the TAP system, while the second would focus on feelings and attitudes towards the system. During the first interview, I asked questions that were related to their experiences with TAP (see below).

1. Tell me about your experiences as a teacher and/or evaluator.
    a. How long have you been teaching?
    b. What do you teach?
    c. How did you get into teaching?
    d. How long have you been at Desert?
2. Tell me about your role in the TAP system.
3. Tell me about your experiences with TAP.
    a. Describe your typical experience with observations/evaluations/conferences/etc.
4. Describe how TAP looks in your classroom. How does it affect your teaching?

a. (For evaluators) Describe how TAP looks in teachers' classrooms. How does it affect their teaching?

5. Tell me about your SKR scores.
    a. Are these consistent over time?
    b. Are they reflective of your teaching abilities?
    c. (For evaluators) Tell me about teachers' TAP scores. Are they consistent over time? Are they reflective of their teaching abilities?
6. Tell me about your SGP (i.e., value-added) scores.
    a. Are these consistent?
    b. Are they reflective of your teaching abilities?
    c. How do these affect your teaching?
    d. (For evaluators) Tell me about teachers' SGP scores. Are they consistent over time? Are they reflective of their teaching abilities?
7. Are the SGP and SKR similar, or do you see discrepancies?
    a. Which one is a better indicator of your teaching abilities?
8. (If the participant was at Desert before TAP) Describe the transition into TAP at Desert.
    a. Describe the school culture.
    b. Describe the pros/cons of the TAP implementation from your perspective.

During the second interview, I asked questions related to the fairness of the system, as well as the participants' overall satisfaction with TAP (see below).

1. Describe TAP in terms of fairness. Is it a fair evaluation system? Why/why not?

2. Describe how (or if) TAP motivates you to be a better teacher.
3. How does TAP affect your relationships with your colleagues.
4. Describe TAP in terms of trust (e.g., between the teachers and evaluators).
5. What happens if a teacher and evaluator disagree on a score?
6. Overall, what do you like about TAP?
7. Overall, what would you like to change about TAP?

I must point out, however, that I also allowed the participants to talk about points of interest that came up naturally during our conversations. As such, each interview was meaningful to the study, yet also unique. I recorded each interview and then manually transcribed each one, using HyperTranscribe software, which yielded approximately 350 pages of transcript data. During this process, I transcribed the interviews verbatim, but I also noted nonverbal cues, such as pauses, sighs, laughs, and the like.

Data Analysis. Data analysis was both an ongoing and reflexive process. I began the analytic process at the very beginning stages of data collection, while continuing to read the literature regarding governmentality and technologies of governance (Foucault, 1991; Foucault, 1977) and neoliberalism as a governing strategy (Rose, 1999). Also, during the data collection and transcription stages, I took detailed analytic memos, allowing me to explicitly track my thinking, questioning, and theorizing (Saldaña, 2006). Too see a sample of my analytic memos, refer to Appendix B.

After transcribing, I conducted two rounds of coding. For the first round of coding I used HyperReseach software, which does/aids with XYZ. In an effort to take stock of what I was dealing with, I began by applying descriptive codes to all of the transcript data, using Saldana's (2013) first-level, open coding. This initially yielded 41 codes (see

Appendix C). As I did this, I took frequent steps back to compare similar codes from different transcripts.

While descriptive coding was helpful in my initial step in understanding the scope of the data, as well as helping me to see similarities and differences between the participants, I found that these comparisons also stripped the excerpts from their contexts. From an epistemological stance, I found it difficult to make meaning from viewing the excerpts in a vacuum-like state. As such, for my second round of coding, I analyzed the data by case—each case consisting of the transcripts of a single participant. During this round, I used Scrivener software, which is a writing software package rather than a traditional a CAQDAS software program. However, I left in the codes from the first cycle on the data, as this allowed me to remain consistent in how I looked at various topics. Also by doing this stage of analysis by case instead of the transcripts as a whole, it allowed me make comparisons between the cases without making generalized assumptions about the group as a whole. Then I performed cross-sectional and categorical indexing to build on the individual cases by noting similarities, contradictions, and other patterns (Mason, 2002). This interpretive process led me to draw conclusions about how the teachers have taken up and embodied a neoliberal discourse, thus defining themselves and their teaching quality in terms of their market value, and disciplining themselves as acceptable teachers.

Of critical importance here, is that, given my theoretical and epistemological positions, I must say that I did not treat these transcript data as a representation of some valid truth or "descriptive, realist tales that would produce a generalizable set of variables in teachers' practices" (Davies & Bansel, 2007, p. 257). Rather I used these data as a

means to makes sense of the potential effects of a neoliberal discourse about teachers and teacher quality as seen through the descriptions and stories of these particular teachers and their evaluators. In other words, I did not attempt to use the transcripts as a means to understand which participants were closer to some external truth (regardless of discrepancies among the responses). Rather, I took each person's transcript as his/her own truth, and building from that assumption, I applied a theoretical lens to draw connections between the present policy landscape, the local context, and the teachers' explanations of themselves and their experiences.

Tying the Two Approaches Together. While I called upon Foucault's work as a theoretical and analytical framework, it is important for me to mention that I do not claim to represent a "true" version of what one might want to call a Foulcauldian study. To do so would be impossible, as Foucault adamantly refrained from categorizing himself or standardizing his methods of doing discourse analysis; instead his "work [was] rather unsystematic," (Andersen, 2003). Instead, I have called on his foundational work, as well as the work of others who have applied and built upon his theory and concepts to organize and make sense of my own study.

Researcher's Role, Responsibilities, and Trustworthiness

First and foremost, I am working from the assumption that nobody, including me as the researcher, is capable of getting outside of discourse (Rabinow & Rose, 2003). In other words, discourse is not a representation of some concrete, physical thing that can be understood and analyzed from an external focal point. Rather, discourse is the imaginable and works to construct the reality in which we all inhabit. As such, I must recognize that through every stage of this study, from its design to its realization, I have developed

interpretations in accordance to my own subjectivity. I do not believe it is possible to be fully aware of my own subjectivities because I am not fixed by particular identities (e.g., researcher, teacher, etc.) that would help me to realize such; rather, my subjectivity is constantly shifting and being (re)negotiated through each interaction and experience I encounter. As such, I have made concerted efforts to be *more* aware of my subjectivities, rather than to deny that they exist or believe that I can get to a point of value-free judgment. This is particularly important as I attempt to build researcher trustworthiness on the account of this study's readers—e.g., . I will do this in two ways. First, I will briefly present my own story as it relates to this study. Then I will explain my plan to make my analytical processes and decisions as transparent as possible, as to allow the readers to not only gain trust in me as the analyst, but also to build their own inferences and conclusions as they see fit.

"Foucault himself starts with the questions: 'What can I do? What do I know? What am I?' These questions are not asked of a unified knowing subject but of a constructed 'I.'" (Brown, 2000, p. 26). The birth of this study rests on the intersection of two personal experiences of mine—my former experiences as an English teacher and my experiences as a research assistant for the Arizona Ready-for-Rigor Project. My experiences as a teacher were what drew me to studying teachers and teacher experiences in the first place. But in this latter position as a research assistant, where I was responsible for delivering presentations regarding the value-added component of the evaluation system (i.e., I explained the model, the calculations, the reasons that growth measures were "better" than previously used status measures, and I answered questions), was what caused me to question teachers as subjects within the evaluation system.

As a research assistant, I was also responsible for calculating the teacher-level value-added scores for the districts. During these experiences I came to the realization that teachers receive information from several sources (including, like me, researchers from universities) that may influence the way in which they accept or deny certain practices (e.g., being measured by value-added models). Recognizing my own contributions to this information delivery, I started to think about how teachers' knowledge about themselves, their practice, and their peers is a complicated network of processes—one that is impossible to reduce to one power-wielding institution or force. I explain this because I hope to give some insight into 1) what led me to use the theoretical approach that I did, and 2) how I am positioned within and inescapable from this work.

Relatedly, I should mention reliability and validity here. As I do not intend for my work to be generalizable beyond the local context of Desert Middle School, reliability—or the probability of the same procedures yielding similar results—is not an appropriate goal. Similarly, validity—or the extent to which my analysis has captured the truth—is also an inappropriate goal. As I have discussed, I am working from the assumption that language and reality are both contextual and negotiable based on time, place, interactions, and the like. That said, I feel the responsibility to make my work and decision-making processes as transparent as possible. Thus, I have made available, in various ways, examples of the data included in the analysis, tables and figures that demonstrate connections between the data, analysis, and findings, and analytic memos detailing, explicitly, my thought processes throughout the stages of data collection, coding, and analysis. This should allow the reader to get an idea of how my thinking was shaped along the way and what evidence I used to justify my thoughts and decisions (see

Appendix B).

Related, a mention of reliability and validity in qualitative research would be appropriate here. While I worked from a naturalistic paradigm and thus relied on the assumption that researchers are never capable of (or intending to) positing truth statements or value-free judgments, it would be remiss to altogether ignore the need to establish credibility and dependability for readers. Rather than attempt to make objective inferences, I worked to draw explicit links between the data and the phenomenon in question (i.e., the problematization of teachers and teacher quality). Above, I discussed the transparency of my analytic process, which can be used to establish what Guba (1981) called an "audit trail" to help build dependability. Similarly, Guba also recommended that practicing reflexivity could help build confirmability by making explicit the epistemological assumptions upon which both questions and inferences were made along the analytic trail (see also Ruby, 1980). This was something I attempted to make visible both in my theoretical framing (see beginning of this chapter) and analysis stages, paying particular attention to shifts in thought along the way.

Limitations of the Study

As with every study, this study comes with its limitations. To some, the most apparent might be the inability to draw generalizable conclusions from these findings. While this would likely be a drawback for traditional policy analysts and evaluators, my intentions for the study have a different outcome goal. Instead of attempting to use Desert Middle School as a microcosm to make grand inferences about the value of particular evaluation practices or instruments, I aim to challenge the way in which we think about knowledge and knowledge production as it relates to teachers and teacher quality. In

other words, I am interested in how teachers are made knowable, or objects of knowledge, and how that knowledge is linked with defining teachers and teacher quality in narrow ways. For, when we solely rely on large-scale, generalizable studies, we miss an opportunity to understand how such approaches discursively affect individual people and practices.

That said, there should also be a consideration for the possibility of naturalistic generalizability (Stake & Trumbull, 1982). Naturalistic generalization refers to the potential of the applicability of one study in one context to another similar context. Stake and Trumbull argue that education practitioners can learn new knowledge vicariously through the reporting of experiences by a researcher. To this end, I aim to provide ample information about the study context, as well as evidence from my observations so as to allow readers to draw from my study that which is most relevant to their own experiences and build upon their own knowledge.

Another limitation of this study is the absence of considering factors such as race, class, gender, or other social factors that might be at play at Desert Middle School. While these factors should be further explored in future studies, the analytical tools with which I worked limited my ability to include these factors in my analysis. This is something of particular interest to me as I continue my work to understand the discursive nature of education policies generally, and teacher evaluation policies specifically. While my study might not address these factors directly, the findings of this study have implicit implications for such factors. I will discuss this further in Chapter 6.

Challenges Faced and Lessons Learned

The first and most profound challenge I faced happened at the beginning stages of

data collection. Fortunately, this challenge shifted the direction of my dissertation, which I believe led to a much richer and more thoughtful study. As it was, I had planned to discuss teachers' experiences under a comprehensive evaluation system, with a particular interest in value-added models (VAMs). Having had some experience in writing about and researching VAMs prior to the study, I (now) realize that I may have had some pre-conceived ideas about how teachers may have negative reactions to such instruments. With this in mind, and a critical discourse analysis (CDA) framework in hand, I was surprised when I heard the first few participants' positive reactions to VAMs.

While their approval of such practices was of certain interest, my own surprised reaction to this was what led to a new theoretical approach. I began to pay attention to pieces of their responses that I may have otherwise overlooked. In doing so, I started to notice contradictions beneath the surface of their responses, and I started linking these contradictions to the policies, practices, and instruments that may have been contributing to their ways of thinking about themselves and their practice, quality, and worth. This was when Michel Foucault's work in governmentality and Nikolas Rose's work in neoliberalism and numbers began to shape my conceptualization of the project as a whole. As such, I stopped looking at the contradictions as units of analysis, and I shifted my focus to start trying to make sense of the conditions that must be present to make such contradictions even possible. It was at this point that I realized that policy research could benefit from approaching the topic from a different angle. This became a driving motivation for this dissertation.

As such, going into the project, I expected for power to be confined to those in leadership—the ones doing the evaluating (which I also believed could be linked to grand

narratives at a macro policy level). But what I ended up finding was that power was more allusive than that. Teachers were behaving in certain ways not because the principal was forcing them to do anything, but rather, because the teachers and the evaluators had taken up a particular discourse that defined teachers and teacher quality in terms of a market value. In doing so, this rationality became the way in which they were able to make themselves and their practice into objects of knowledge. By knowing themselves in this way, they could act accordingly. Simultaneously, this way of thinking about themselves and each other created a common system and mission to which the school could function as a team, or an enterprise where teachers conducted themselves as responsibilized entrepreneurs of themselves (Brown, 2003; Rose et al., 2006).

CHAPTER 4

Part 1: Analysis of Policies, Practices, and Instruments

For this part of the analysis, I collected and analyzed policy documents that directly relate to the teacher evaluation system used at Desert Middle School, including official policies about teacher evaluation (i.e., federal, state, and local), official White House and US Department of Education press releases, speeches, and other documents related to the policies, and promotional materials related to TAP specifically (as available on TAPsystem.org). This helped me understand how teachers have been discursively positioned as the problem in need of being solved, as well as how evaluation policies and practices work to define teachers and teacher quality as problems in particular ways. Then I narrowed in on the policies and practices at one school and analyzed materials related specifically to the local context of the study. This helped me understand how one Arizona middle school has utilized practices and instruments in order to manage the conduct of teachers.

The Problematization of Teachers and Teacher Quality

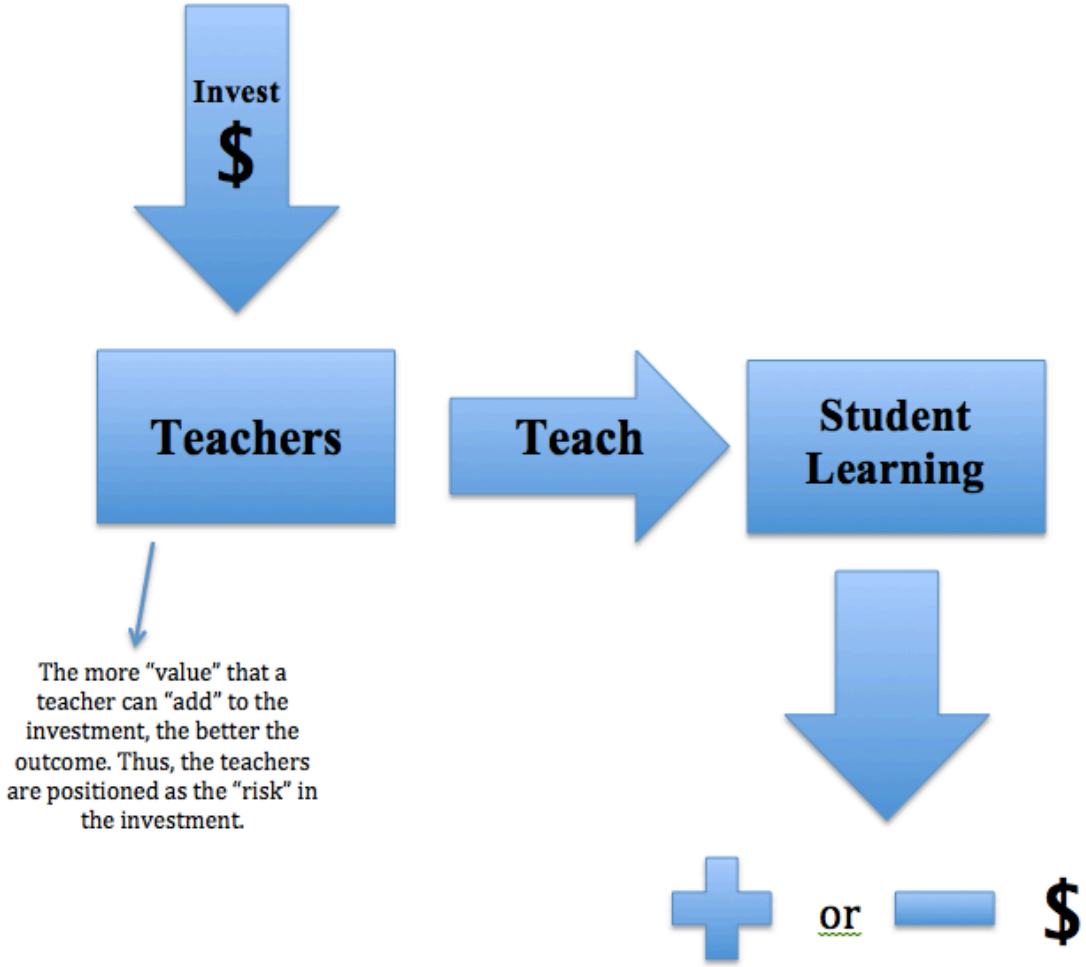
Policies and practices are developed to solve some problem. However, the problem itself is constituted, or defined, by the very policy/practice/tool aimed at solving it (Bacchi, 2000). According to Rabinow and Rose (2013), "to analyze problematizations is not to reveal a hidden and suppressed contradiction: it is to address that which has already become problematic," (p. 13). Since teacher evaluation systems serve the purpose of differentiating teacher effectiveness so as to get rid of "ineffective" teachers, the "problem" to be solved is that of teachers. Given that the current neoliberal discourse defines all aspects of education, including teachers and teacher quality, in terms of their

market value (Peters, 2009), the problem that teacher evaluations attempt to solve is also constituted in terms of market value. In other words, teachers' effectiveness has to be defined and valued based on how it contributes to the economic market. Also, the tools and practices used to carry out such policies also work to define problems. Using this as a framework, I sought to understand 1) how teachers have been positioned as the problem, or problematized in education policy; and 2) how evaluative practices and tools have defined teacher quality by managing teacher conduct in particular ways. As per Rabinow and Rose (2013), I am not attempting to seek the "real" problem in education, but rather to understand how teachers have been constructed as the problem in need of solutions (e.g., stricter evaluation systems), as well as how teacher evaluation policies and practices work recursively to constitute teachers and teacher quality as particular types of problems.

Teachers as Risky Subjects. Generally speaking, a free market compels a certain level of risk. Investors risk money in hopes of making more, but not without the chance of losing it. Such risk, then, calls for risk management in order to (hopefully) minimize the potential risk. As teachers have been re-conceptualized as market-based subjects, they too present some level of risk to an investment. Dean (1999) wrote that by "calling into question some aspect of the [teachers's] 'conduct of conduct'" (p. 27), teachers have been positioned as "risky" subjects in need of being managed, controlled, and disciplined. See Figure 1 for a conceptualization of teachers as a "risk" in education investment.

Figure 1

Conceptualization of Teachers as Part of an Education Investment



The concept of teachers as risky was a consistent trope in the data regarding discussions about teacher evaluation policies (e.g., speeches, press releases, promotional materials). Based on my interpretation of the data, I suggest that this was done it two ways. First, the need for evaluating teachers was couched as a way to ward off U.S. economic failure, as exemplified in President Obama and Secretary Duncan's various speeches. In one of his speeches promoting RttT, President Obama stated that "Countries that out-educate us today will out-compete us tomorrow, and I refuse to let that happen on my watch," (Remarks by the President on Race to the Top at Graham Road

Elementary School, 2010). Similarly, in one of the earliest speeches about RttT, Secretary Duncan stated the following: "[the President] understands that education is the foundation of our economic strategy and the only sure path to long-term economic strength," (Arne Duncan to NEA, 2009).

In both of these instances, education was directly linked to national prosperity in relation to a global economy, and in fact named as the *only* way for the country to succeed in a competitive world. This theme, as well as the link between teaching and "national security" ran throughout the texts. Secretary Duncan went as far as telling Baltimore County teachers "that teachers are the heart and soul of our education system--and that our success as a country is entirely dependent on your success as a teacher," (Duncan to Baltimore County Teachers, 2012). In this example, Secretary Duncan posits that teachers are *entirely* responsible for the economic success of the country. In so doing, he not only dismisses every other factor that may or may not contribute to the success of the country, but he also builds a foundational rationale for any method to keep teachers behaving up to expectation (e.g., rigorous teacher evaluation methods).

Similarly, individual student economic success was also directly linked to teacher quality, as exemplified in President Obama's 2012 State of the Union Address:

> At a time when other countries are doubling down on education, tight budgets have forced states to lay off thousands of teachers. We know a good teacher can increase the lifetime income of a classroom by over $250,000. A great teacher can offer an escape from poverty to the child who dreams beyond his circumstance. Every person in this chamber can point to a teacher who changed the trajectory of their lives. Most teachers work tirelessly, with modest pay, sometimes digging into their

> own pocket for school supplies -- just to make a difference. Teachers matter. So instead of bashing them, or defending the status quo, let's offer schools a deal. Give them the resources to keep good teachers on the job, and reward the best ones. And in return, grant schools flexibility: to teach with creativity and passion; to stop teaching to the test; and to replace teachers who just aren't helping kids learn. That's a bargain worth making (The White House, 2012).

In the excerpt above, not only did President Obama suggest that students' earnings could increase because of a good teacher, but that a student might be left in poverty if he/she does not have a great teacher. In these terms, this is a tremendous risk to consider—for it is being positioned as a choice between a life of prosperity or a life of poverty for each student. Here, "teachers matter" in terms of how they impact the future earnings of students, and for that reason, the country should invest in such a "bargain" to get rid of teachers who are not productive in this way.

Teachers were also positioned as risky subjects by way of threat. There was a consistent theme across the policy discussion data that bad teachers have been left in classrooms across the country, plaguing the system and threatening students at every turn. For example, Lowell Milken, the founder and CEO of TAP, made the following comments in a speech about the need for systems like TAP:

> We know from research that, aside from home and family, the single most important factor driving student performance is the quality of the teacher in the classroom. The difference between an effective and ineffective teacher can be a full grade level of student achievement in a single year. Now, based on these facts, you would think that every effort would be made within the K-12 system to implement

> a structure that would attract large numbers of talented people to teaching, and then create an environment in which they would thrive. Sadly, however, this is not the case. The fact is that none of the hundreds of costly school-reform efforts over the past decades have had the scope, force and focus to attract high-caliber talent to the teaching profession, and then reward and motivate the talent to stay. That is a primary reason why, more than 50 years after Brown v. Board of Education, well over 50 percent of all African-American and Hispanic fourth-grade students cannot read and barely one-third of fourth-, eighth- and twelfth-grade students in our nation reach NAEP proficiency levels in reading or math, (Milken, 2005).

Above, Milken directly linked teacher quality and the reading proficiency levels of African American and Hispanic students, thus blaming teachers for low test scores. Again, teachers were presented as a high risk to students, and teachers' conduct was foregrounded as the problem with student success. This ties back to the concept of teachers as an investment that can either add value (as in President Obama's speech) or detract value (as in Milken's speech). Accordingly, tools for disciplining teachers and governing their behavior to be more aligned with such desired forms of productivity and value were rationalized via neoliberal logic.

In contemporary society, explicit force has been replaced with apparatuses that encourage teachers to manage themselves in relation to risk (Saul, 2005). McWilliams and Jones (2005) argued, "in terms of contemporary working life, risk as a moral climate offers new ways of being properly professional, one of which is alertness to potential dangers and greater attention to the work of minimising the possibility of something going wrong," (p. 110). As such, there manifests a need to minimize the risk and develop

tools for managing the potentially risky subjects.

Managing Risk through Teacher Evaluation Policies and Practices

The positioning of teachers as risky also implies a need for policies and practices to help manage such risk. While the need for such practices was rationalized by the "risk," the solution, or management of that risk, presents another set of problems. Risk management entails creating a system where teachers "question their own conduct, to watch over and give shape to it, and to shape themselves as ethical subjects"(Foucault, 1985, p. 13). Techniques of governance are needed to carry out such a task. Practices and tools, in this sense, actually have an effect on teachers' behavior. In other words, the practices and tools produce particular types of teachers, rather than capture "quality" as an independent construct. For this part of the analysis, I will discuss the practices and instruments designed to manage the teachers' conduct. In Chapter 5 I will discuss the effects this has had on teachers and evaluators at Desert Middle School specifically.

As per Rabinow & Rose (2003), "Foucault uses the word *apparatus* to mean a device oriented to produce something – a machinic contraption whose purpose in this case is control and management of certain characteristics of a population," (p. 10). The positioning of teachers as a problem in need of being disciplined has called for apparatuses to be developed and implemented in order to keep teachers performing in a desired manner. Various instruments, tools, and practices have not only been developed to discipline teachers to behave in certain ways, but they also encourage teachers to discipline, or govern, themselves. For this part of the analysis, I have used official policy documents that specifically relate to the local context of the study (i.e., Arizona's RttT application, Arizona Ready-for-Rigor application, Arizona Framework for Measuring

Teacher Effectiveness (2011), and Arizona's ESEA Flexibility application) as well as local evaluation practice protocols (i.e., TAP evaluation system practices and instruments) in order to show how tools have been developed to govern teachers' conduct. I sought to understand how the practices and tools function together to create "technologies of governance" or "the intellectual and practical instruments and devices enjoined upon human beings to shape and guide their ways of 'being human'," (Rainbow & Rose, 2003, p. 16).

Numericization and Objectification of Teachers

Numbers provide a mechanism for measuring the health, or state, of populations and other social matters (e.g., poverty, economy, health, etc.) (Rose, 1991). Education has been no exception. The rise of globalization has been accompanied by a call for a market-based way of thinking about teachers, students, teaching, and learning. "Accountability" has dominated the discourse on education for the past thirty years, thus leading to a need for numericizing aspects of education. Human judgment has been replaced with objectification, for "numbers are part of the techniques of objectivity that establish what it is for a decision to be 'disinterested'," (Rose, 1999, p. 199).

The evaluative instruments and practices that have been used at Desert Middle School add to the numericization of teachers and teaching practices. Most obviously, VAMs, not only quantify student achievement and growth, but they also quantify teaching by attributing that growth to teachers. As per the policies that I collected and analyzed (at the federal, state, and local levels), the use of student growth in teacher evaluations is held as a priority. As stated in RttT: "[States must] establish clear approaches to measuring student growth (as defined in this notice) and measure it for

each individual student," (section D(2)(i); p. 34, RttT Application, 2009). As I discussed in the literature review, value-added models (VAMs) and student growth percentile (SGP) models are the most popular methods of doing so across the country. Though these two models differ in their statistical properties and functions, for the purposes of this analysis, they function in similar enough ways to discuss them as a single apparatus. As such, I will refer to them as VAMs from here forward, which is defined as a statistical model that intends to measure the effects of teaching on student learning via standardized achievement tests over time.

VAMs, regardless of their statistical capabilities (or limits), serve the function of objectifying the relationship between teaching and learning. In doing so, teaching and learning are defined in terms of what and how VAMs quantify such subjects. Thus teacher quality is reduced to a number that is subject to those *who* chose to measure it and *how* they chose (or were able) to measure it (Rose, 1999). VAMs turn teacher quality into objects of knowledge that are then made subject to measurement, comparison, and evaluation.

Similarly, while observations (i.e., classroom observations and artifact/lesson plan submissions) are sometimes referred to as the more qualitative measure of teacher performance/practice, these are also numericized by the use of rubrics. In order to meet the requirements of the Arizona Framework for Measuring Teacher Effectiveness (2011) and other relevant policies, teachers must be evaluated on a numerical system (i.e., the "Teacher Performance" component of the total teacher evaluation calculation must be between 50% and 67% and the rest must come from the student growth component). As such, all aspects to be included in the evaluation calculation must be turned into numbers,

or objective units of knowledge.

This process also turns teachers (and teacher quality) into knowable subjects. Aspects of their practice, performance, and the like, are made objects of knowledge that can be acted upon and managed. They are measured, compared, and ranked based on the numbers assigned to such aspects. Their worth is constituted by the way in which numbers can be applied to their practice. Again, though, the way in which numbers are used to define teachers and teacher quality is always subject to *how*, *why*, and *by whom* such numbers are determined (Rose, 1999). This creates certain types of teachers and simultaneously eliminates other ways of thinking about teachers, or, perhaps more importantly, it limits different ways of *being* a teacher. Numbers, essentially, make aspects of teaching visible in ways that were previously impossible. This visibility also subjects teachers to the technology of surveillance.

Surveillance

Hierarchical surveillance, or the technique of making subjects visible and observable, is a foundational component of creating governable persons (Foulcault, 1977). Teachers, as part of comprehensive evaluation systems, are subject to constant surveillance. This is carried out explicitly via classroom observations (i.e., formal and informal), as well as through artifact and lesson plan submissions. Teachers and their practice are made further visible through pre- and post-conferences as teachers are expected to make visible their thinking that goes into their lesson planning, which is done verbally as well as written by way of conference forms and checklists. One DMS evaluator called this the "consciously competent" teacher—one who is able to explain his/her decisions about classroom practices (further explained in Chapter 5). Once made

visible, the teachers' practices and/or thinking are subject to examination and thus evaluation.

Surveillance disciplines teachers to perform in specific ways because 1) their worth (and thus their job and/or pay) are valued based on how they perform under surveillance, and 2) they never know when they might be surveilled. Foucault (1977) wrote: "The exercise of discipline presupposes a mechanism that coerces by means of observation; an apparatus in which the techniques that make it possible to see induce effects of power, and in which, conversely, the means of coercion make those on whom they are applied clearly visible," (p. 97-98). The more visible the teachers and their practices are, the more subject to audit they are and thus more governable.

At DMS, observations serve as a key component to their evaluation system. Teachers are observed in practice via formal and informal observations by a principal or master/mentor teacher. They have four formal observations, two of which are scheduled ahead of time with the teacher, and two of which are surprise visits. All four formal observations take place for an entire lesson (45 minutes for most, and 90 minutes for block classes). During the observation, the evaluators are trained to "capture evidence" by following the guidelines below (as per the TAP system protocols):

1. Time: Capture the length of different segments of the lesson.
2. Abbreviate: It's tough to get down everything the teacher says or does, so , when possible, abbreviate. After the lesson, review your notes and write out what you abbreviated.
3. Verbatim: Capture verbatim dialogue when possible. Nothing is better than direct quotes of what the teacher and/or students say. Use T for

teacher and S for student.

4. Paraphrase: Use parentheses to indicate that you are paraphrasing, so when you go back through your notes you know what is paraphrased and what it verbatim.
5. Q & F: After you finish, go through all questions and feedback.
6. Upfront Summary: After you finish, go through your evidence and write a brief summary of the lesson.
7. Label: Begin to categorize your notes by labeling evidence for various indicators on the rubric.
8. Lesson Analysis: Identify the lesson's primary objective and its sub-objectives.
9. Circulate: Circulate as necessary to collect evidence from the teacher, students, and student work.

The observer has also been trained to fill out a "Teacher Observation Report Template" that requires she/he to rate the teacher on the 19 TAP rubric indicators using a scale of one to five (a form of numericization).

Teachers are also subjected to informal observations, or "walk-throughs" where either a mentor, master, or administrator will observe the teacher for approximately five minutes. These are unannounced and not restricted to any time period or quantity. During this time, the observer is to pay attention to one element of the teaching (also aligned to the TAP rubric) and provide feedback to the teacher in an informal manner (i.e., there is no official post-conference, but there might be an informal follow-up).

Another form of observation that is part of TAP at DMS is the observation of

written lesson plans. As part of the formal observations (i.e., classroom visits), teachers also have to submit written lesson plans that are assessed based on certain TAP criteria. These evaluated lesson plans become part of the teacher's overall evaluation (i.e., part of the SKR score). This is another way that teachers are to make their practice visible, assessable, and evaluate-able.

Teachers and their evaluators convene for conferences before and after announced formal observations. They also meet for post-conferences following unannounced formal observations. During this time, the evaluator conducts the conference based on a specified TAP model that includes pre-determined reflection questions. The evaluator has been trained to guide the teacher through a series of self-reflection questions that encourage the teacher to think about his/her instructional decisions. Examples of pre-conference reflection questions are:

1. What are the pre-requisite skills needed for students to be successful?
2. What changes or adjustments will you need to make if students do not show evidence of mastery of the sub-objectives?
3. How will you know students have mastered the objectives?
4. Is there anything you want me to be aware of before the observation?
5. How will you differentiate your instruction to address various learning styles?

During the post-conference, teachers and their evaluators compare their scores (the teachers score themselves on the rubric as well). After presenting the evidence from the lesson, the evaluator is to highlight an area for reinforcement (performed well) and an area for refinement (needs improvement). Like the pre-conference, evaluators have been trained on how to conduct such sessions, following specific guidelines and pre-

determined questions. Based on the protocols, as well as discussions with the participants about this practice, there appeared to be an underlying goal for the teacher and the evaluator to eventually synchronize their ratings based on normalized judgments (Foucault, 1977; 1984).

Normalizing Judgments

The act of evaluating teachers rests on the technique of "normalizing judgments" (Foucault, 1977; 1984). In order to measure, evaluate, and rank teachers, there must also be a way of normalizing teacher behaviors, or the process of developing a "normal" way of behaving so that teachers can be compared against such a norm. Standards-based education, in general, functions in such a way as well. "In a sense, the power of normalization imposes homogeneity; but it individualizes by making it possible to measure gaps, to determine levels, to fix specialities and to render the differences useful by fitting them one to another," (Foucault, 1977, p. 103).

TAP, as well as other evaluation systems that are similarly fashioned by RttT and other relevant policies, has been built on normalized judgments. Again, in order to remove human judgment as much as possible, evaluators must have tools that allow them to make similar judgments to one another (e.g., inter-rater reliability). Instruments of numericization (e.g., VAMs and rubrics) allow for such judgments to be made. The rubrics have a set of "normative criteria" that are accessible to all teachers and evaluators, making it possible for different competent observers to make similar conclusions about the teachers' quality (Ransom, 1997, p. 171).

One way of ensuring normalized judgments is by use of rubrics. Rubrics are a fundamental component to the evaluation system at DMS, which is also a common

practice in other districts and states. Rubrics are designed to capture particular aspects of teacher quality and translate that quality into some numerical value that allows for the measurement, comparison, and evaluation of said quality. Rubrics also serve as a tool to change teacher practice. According to teachers and evaluators at Desert, the rubric is often used in weekly cluster meetings as professional development. Mentor and master teachers develop lessons based on the rubric indicators, and then they complete field tests of the lessons with their own students. If they deem the lesson successful, they teach the lesson to their colleagues at the cluster meeting. The lessons are designed to provide teachers with targeted lessons that should help them increase their scores on the TAP rubric during their observation evaluations.

Another method of normalizing judgments was through the evaluator training course and certification. During this course, the evaluators practiced evaluating teachers by watching a video of a lesson, collecting data, scoring the lesson via an observation rubric, and then discussing their evidence/scores with the instructor and class. During this time, the trainer would reveal the "correct" (i.e., the official TAP evaluators'/trainers' scores) to the group. There was little allowance for discussion, as it was made clear that there were correct answers and wrong answers. At one point, the evaluators realized that if they scored everything with threes (on a one to five scale), then that would keep them within the acceptable range of earning their certification (i.e., to pass the certification exam, evaluators were required to evaluate a lesson and be within one point of the correct scores on each of the TAP indicators). Since it was made clear by the evaluator trainer that it was rare for a teacher to earn a one or a five on an indicator (and evaluators were encouraged to stay away from these), then a score of three, almost every time, would be a

safe way to be within the acceptable range. As such, the evaluators-in-training (including myself) responded by scoring with mostly threes, regardless of opinion.

Similarly, evaluators are trained on how to conduct pre- and post-conferences with teachers. The training materials include guidelines, forms, and sample pre- and post-conference question prompts. During the training, evaluators were required to develop mock conferences using the materials. Then the evaluators practiced conducting the conferences with other trainees. There was an expectation that evaluators stick to the script (as per the guidelines, forms, and prompts) and base their recommendations for reinforcement and refinement areas on the rubric scores and collected evidence (i.e., evaluators were encouraged to script lessons, or write down word-for-word the teacher and student dialogue in the lesson).

This training process has two key effects that should be noted. First, it creates a limited scope within which the evaluators can think about teacher quality. With a focus strictly on that which TAP rubrics require, then other possibilities of being an effective teacher might be missed or devalued. This was also evident in the teachers' interview when they had to make choices about what was good for their students versus what the TAP rubric expected of them (more on this in Chapter 5). As such, while creating normalized judgments might be intended to create fairness across evaluators, context, and time, it also leads to a production of certain types of teachers. This normalized way of thinking about teachers leads to a "regime of truth" (Foucault, 1980) that constitutes proper behavior (McWilliams & Jones, 2005). According to Dean (1999), "We govern ourselves according to what we take to be true about who we are, what aspects of our existence should be worked upon, how, with what means and to what ends. We thus

govern others and ourselves according to various truths about our existence and nature as human beings," (p. 18). In other words, the rubric provides specific types of "normal" conduct to which teachers are expected to adhere. Once this is taken up as the way to be "normal," the teachers also embody this discourse and begin to not only judge and qualify themselves against such standards, but they will also adjust their behaviors accordingly. I will discuss more evidence of this phenomenon in the teacher interview data in Chapter 5.

Examination

Examination is made possible via the technologies of surveillance and normalizing judgments (Foucault, 1977). Making something visible (surveillance), then developing standards against which to make comparisons (normalizing judgments), allows for the examination to take place. The examination is the process by which something is assessed and qualified. "It establishes over individuals a visibility through which one differentiates them and judges them," (Foucault, 1977, p. 103). While students have been subjects of explicit, formal examination via standardized assessments for decades, this process has also made it possible for teachers and other stakeholders to be subjected to examination as well (Graham & Neu, 2004, p. 311). The TAP practices and instruments that are used at Desert Middle School only expand on this process to make explicit the direct examination of teachers.

VAMs and rubrics attempt to make visible teacher quality; however, they simultaneously constitute teachers and teacher quality in specific and confining ways. The examination is not just about assessing that which exists. Instead, examination produces certain types of individuals by problematizing the construct that the

examination is attempting to capture. As subjects internalize the examination outcomes, they begin to think about themselves relative to the examination instruments and their peers—and so begins the process of the individuals modifying their behaviors in order to fit into the norm (McWilliam, 2002). The TAP rubric stands as the foundation of not only teacher evaluation at DMS, but also the professional development. In this sense, the rubric, though an inanimate object, has a real impact on the ways teachers behave. Similarly, VAMs that rely so heavily on student achievement scores on large-scale standardized achievement tests, also have effects on the way in which teachers behave. This is a parallel finding with other education policy analysts who have looked at teaching to the test, marginalized content areas, and other forms of system gaming (Cawelti, 2006; Darling-Hammond, 2007; Menken, 2006; Smyth, 2008).

The impact of a seemingly mundane object (e.g., rubric) can be further unpacked by looking at a discussion that took place at an official "Forum on ESEA Flexibility" (available on Ed.gov). The forum was open to states who were interested in applying for ESEA Flexibility. It demonstrates how a seemingly simple decision—one of choosing an appropriate observation rubric—can be steeped in a market-based discourse and can have effects on how teachers and teacher quality are constituted by seemingly mundane decisions. The speaker in the excerpt below, Ms. Heyburn, was a policy advisor who worked in the Tennessee Department of Education (DOE) when Tennessee transitioned from the Charlotte Danielson (1996) observation rubric to the TAP rubric. Representatives from the Tennessee DOE were asked to share their experiences with implementing teacher evaluation policies as based on ESEA Flexibility.

Before discussing Ms. Heyburn's comments and justification of rubric choice for

Tennessee, I want to first bring attention to the rubric as an instrument that serves two functions—to "capture", or measure, teacher quality and to influence teacher behavior (to theoretically improve teacher quality). In essence, the tool (rubric) is designed to solve (or improve) the problem of teacher quality. But what is important to point out is that the problem is defined in terms of the tool. Further, the solution to the problem (improved teacher quality) is dictated by the tool as well. In other words, the teacher's quality is a problem if, and only if, the specific rubric used to measure the teacher's quality deems it a problem. For example, the TAP rubric has a category for problem solving, which indicates that a teacher at the highest level "implements activities that teach and reinforce three or more of the following problem-solving types: abstraction, categorization, drawing conclusions/justifying solutions, predicting outcomes, observing and experimenting, improving solutions, identifying relevant/irrelevant information, generating ideas, creating and designing" in every lesson. A teacher who focuses on one of these in-depth, or a teacher who does something else entirely for a lesson, is identified as lacking in this area and in need of improvement (i.e., practicing in the way the rubric states as good). In other words, the teacher has a problem with critical thinking; yet this was not necessarily a "problem" that existed prior to the use of the rubric. Since this is the way the teacher's quality (for critical thinking) has been defined, the teacher will modify (or is expected to modify) his/her practice to meet the expectations of the rubric, thus eliminating other possible ways of teaching critical thinking. Below, Ms. Heyburn discusses Tennessee's choice of rubric in response to the question from the audience: "I got the impression that you moved away from Charlotte Danielson's work, and I'm curious, if you did, why did you? And, secondly, what are -- what is the model then for

quality of practice that you're using?:

> Ms. Heyburn: Yeah, I'm happy to take it. So, first, I think, you know, a lot of the rubrics that we looked at are rooted in Charlotte Danielson's work. You know, there is only so many kind of domains of practice that we really as educators all agree to. So one of the things we first noticed was that, you know, planning, instruction, environment, professionalism, you know, which are at the heart of her work are at the heart of, you know, the rubric that we chose and -- and several of the others that we looked at. And so I think it's fair to say that it's not necessarily in the rubric, but oftentimes it's how it's implemented. And so the rubric that we had been using in our existing framework before we changed to this new model [i.e., TAP] was the Charlotte Danielson rubric. And it wasn't that we weren't happy with that. It was just we needed new resources and new ways to implement a somewhat, you know, new and customized system. So the rubric that we're implementing now still looks a lot like Charlotte Danielson. It's streamlined a bit further, and we were able to provide the resources around, again, kind of the video portal and the inter-rater reliability certification that aligned with this specific rubric. So that was part of our choice given our tight timeline was that we needed to look both at the quality instrument and the ability to take it to scale. And this instrument helped us to that end.

Ms. Heyburn's response is relevant to this discussion for the way in which she rationalized Tennessee's choice of rubric, and then the implications this has for defining the problem (i.e., teacher quality). She provides three reasons for such choice: 1) the TAP rubric includes the domains that "we really as educators all agree to"; 2) they needed a

rubric that had more resources that accompanied it; and 3) they were on a strict timeline and needed something that could easily scale up. In this explanation, she minimized the concern for what the rubric might capture and emphasized the state's need for something quick, scalable, and comprehensive in terms of accompanied resources. The implications of this choice, however, are that teachers will behave in particular ways that would be different had there been another rubric or no rubric at all. I will discuss further the way in which teachers have incorporated this way of thinking about themselves and their behaviors in Chapter 5.

Discipline

When combined, the technologies of numericization, surveillance, normalizing judgments, and examination come together to create a system of discipline (see Table 2 for a description of the technologies of governance at Desert Middle School). For one, teachers face real consequences for not conforming to the system. Given that TAP is a performance-based compensation system (PBCS), one of the key uses of the evaluation outcome data is merit pay. Teachers are divided into different pools depending on various characteristics (i.e., career, mentor, master, and hard-to-fill). The teachers' composite evaluation score is ranked among their peers in their same pool, and the money is split respectively. According to the Arizona RttT application, the following outcome uses (i.e., the ways in which evaluation scores are to be used in personnel decisions) have also been required:

1. Developing teachers and principals, including by providing relevant coaching, induction support, and/or professional development;
2. Compensating, promoting, and retaining teachers and principals, including by

providing opportunities for highly effective teachers and principals (both as defined in this notice) to obtain additional compensation and be given additional responsibilities;

3. Whether to grant tenure and/or full certification (where applicable) to teachers and principals using rigorous standards and streamlined, transparent, and fair procedures;
4. Removing ineffective tenured and untenured teachers and principals after they have had ample opportunities to improve, and ensuring that such decisions are made using rigorous standards and streamlined, transparent, and fair procedures. (section D(2)(iv); p. 34, RttT Application, 2009).

Table 2

Technologies of Governance at Desert Middle School

Technology	Function	Practices/ Instruments
Numericization	The process of turning matters into numbers—making teachers knowable as objects of knowledge (Rose, 1999)	Rubrics Value-added models
Surveillance	The making of teachers, as well as teachers' practices and attributes visible both explicitly (e.g., observations) and implicitly (e.g., making their thinking visible) (Foucault, 1977)	Observations Lesson plan submission Pre- and post-conferences Data dashboards
Normalizing Judgments	The setting of a standard or normal way of making judgments about teachers so that comparisons can be made about them (Foucault, 1977)	Rubrics Evaluator forms/guidelines

		Evaluator training handbook Evaluator conference prompts Evaluator training course
Examination	The combination of surveillance and normalized judgments—teachers are made visible and then examined based on normalized judgments (Foucault, 1977)	Rubrics Value-added models Observations Lesson plan submission Pre- and post-conferences Data dashboards
Discipline	The combination of numericization, surveillance, normalizing judgements, and examination that disciplines teachers to conduct themselves in desired ways (Foucault, 1975)	Evaluation outcome use (i.e., personnel decisions as based on evaluations) Cluster meetings/training Self-reflections Pre- and post-conferences Observations Rubrics Value-added models

As mentioned, explicit force has been replaced by techniques of self-discipline. While teachers might lose the opportunity to make extra money via merit pay or face the

loss of tenure or even their jobs, "[e]xaminations, besides producing these external disciplinary effects, also encourage the internalization of disciplining activities," (Rabinow, 1984, p. 19). In other words, it is not only the explicit threat of consequences, such as termination, that produces disciplinary power, but rather, the way that individuals discipline themselves in light of the disciplinary tools and constructed group norms (McWilliam & Jones, 2005). One example of self-discipline is the way in which teachers are encouraged to self-reflect on their practices. Before meeting with their evaluators, they are required to score themselves on the TAP rubric, and they are to determine what they think their areas of reinforcement and refinement should be. These become part of the conversation with the evaluators, but, according to the teachers and evaluators at Desert Middle School, the evaluators usually have the last say given their experience and "evidence" at hand to make their claims. This reinforces the idea that there is a correct, or normed, way to gauge one's quality.

I will discuss in Chapter 5 how the teachers at Desert Middle School specifically have taken up this way of thinking about themselves and their peers. In so doing, they have created a binary upon which to judge each other that categorizes teachers as acceptable or unacceptable—all of which have been constructed based on the disciplinary tools and technologies that I have discussed herein. This dichotomous view of teachers has created a system where teachers behave in particular ways so to avoid the label of being unacceptable.

Audit-able Teachers

Neoliberalism as a form of governmentality produces a "culture of audit" (Power, 1997) that relies on a system of accountability (Hodkinson, 2008). According to

Foucault's analysis of neoliberalism: "through setting individuals against each other in intensified competitive system of funding with clearly defined measures of success, those individuals are de-individualized and converted into the generic members of an auditable group," (Davies & Bansel, 2010). As teachers are numericized and then subjected to surveillance and examination, they become objects of audit in a society that conceptualizes schooling as a market-based endeavor where learning is a product. Teachers' value is thus conceptualized and constituted in terms of their [in]ability to add value to their products. VAMs and rubric-based evaluation practices make possible these ways of thinking about teachers and their worth in our society. Teachers, as well, take up this discourse and begin to see themselves and compare themselves in such ways (this process is at the heart of Chapter 5 where I talk to teachers about their positions within the TAP evaluation system).

In thinking about audit as a technology of governance, or as a technique of getting people to conduct their behavior in desired ways, then the focus should not be on how audit oppresses individuals, but rather on how such methods produce "responsibilizied and accountable subjects," (Davies & Bansel, 2010, p. 9). For example, TAP teachers are required to self-reflect and self-score before meeting with their evaluators. More implicitly, teachers are under constant surveillance via informal classroom visits, lesson-plan submissions, and data dashboards that allow evaluators to observe classroom test scores at any time. These data can be used at any time, and in many ways unknowingly to the teachers. Graham and Neu (2004) wrote:

> Foucault (1984) refers to this as the 'panoptic modality of power', in that it is impossible to know when or even if the numerical traces will be used. It is this

invisibility of potential users and usages that disciplines participants. Furthermore, resistance to any such use is impossible, because one cannot say who is using what, and why, or when," (p. 311).

As such, teachers are subjected to a constant state of accountability, and more importantly, self-accountability, which results in 'entrepreneurial actors' (Brown, 2003, p. 38).

Conclusion

For this first part of the analysis, I started by collecting and analyzing policy documents and other texts relevant to contemporary teacher evaluation systems. Specifically, I looked at the TAP model, which has been marketed to meet the requirements of RttT, TIF grant, and ESEA Flexibility applications. I first looked at how teachers and teacher quality have been problematized, defined, and constituted by such policies and policy discussions. Then I looked at particular TAP evaluative practices and instruments using policy and TAP documents, field notes and training materials from the TAP evaluator certification course, and TAP instruments (e.g., rubrics, conference preparation forms, etc.). I analyzed these using Foucault's framework for neoliberalism as a governing form, while questioning the use of such practices and instruments in terms of how they work to manage the conduct of teachers, while simultaneously working to produce teachers who manage their own conduct.

I found that not only have teachers been positioned as a problem in need of being fixed, but more importantly, that the "problems" with teachers and teacher quality have been defined by the very policies and practices meant to fix them. In other words, teachers and teacher quality are subjected to why, how, and by whom they can be

measured. As such, particular types of teachers are produced as a result of the policies and practices. TAP rubrics, for example, are designed to measure teacher quality; however, they actually work to create teachers who behave in particular (and limiting) ways. Other practices, such as VAMs, observations, conferences, and personnel decisions based on evaluation outcomes come together to create a culture of audit. Teachers are numercized and then subjected to constant surveillance and examination. In this sense, teachers are turned into objects of knowledge and are disciplined to not only behave in desired ways but also to discipline themselves. Fenwick (2003) wrote: "Practices that render individuals 'knowable' through examination, observation, classification and measurement, control people by making them objects of knowledge" (p. 345).

Neoliberalism as a form of governmentality has created a system whereby education, and thus teachers, are valued in terms of their market value. With an increased focus on globalization (i.e., a global market), schools have been positioned as a means to sustain or increase the United States' economic dominance in the world. In reference to RttT, Arne Duncan consistently made references such as: "It is not just our economic security that is at stake--but our national security as well. A strong military remains our best defense, but a strong education is our best offense" (Moving Forward, Staying Focused: Remarks of Arne Duncan, National Press Club, 2012). To govern by way of neoliberalism and to create self-governable subjects, all aspects of education must be standardized, numericized, and, essentially, monetized. Thus teachers and teacher quality had to also become objects of such a system. The evaluative practices and instruments discussed herein do just that.

Given that "technologies of audit and surveillance, of self-audit and self-

surveillience, are not simply discourses of responsibility and accountability but technologies for the production of responsiblized and accountable subjects" (Davies & Bansel, 2010, p. 10), it is important to examine how such discourses have been taken up by teachers at a school site. Governmentality analysts have found that individuals, as subjects of the discourse, are "made up" (Hacking, 1999) in that they are produced as certain types of subjects (Dean, 1999; Graham & Neu, 2004). In the case here, the teacher who adds/detracts value, or the TAP teacher, are types of teachers who exist only because instruments such as VAMs and TAP rubrics exist. Therefore, for the next part of the study, I sought to understand how teachers at one school have become these types of teachers by analyzing how they describe themselves as knowable subjects within a TAP evaluation system.

CHAPTER 5

Part II: Analysis of Teacher Interviews

In this chapter I attempt to link the evaluation technologies of governance to the way in which teachers have begun to think about themselves, their practice, quality, and worth. To do so, I interviewed a group of teachers and their evaluators at one middle school. In this analysis I demonstrate how teachers have embodied the neoliberal discourse, and in so doing, have begun to define themselves and qualify their practice and worth in terms of market value. Similarly, they have subjected themselves to various techniques of governance, while denouncing other teachers who have chosen not to participate. This justification rests on a binary that the teachers have constructed about what it means to be an acceptable versus an unacceptable Desert Middle School teacher.

Hacking (2004) argued that the way in which we see ourselves and how we make sense of who we are has a real effect on the possibility of who we are able to become. He posed the question: "How is the space of possible and actual action determined not just by physical and social barriers and opportunities, but also by the ways in which we conceptualize and realize who we are and what we may be, in this here and now?" (Hacking, 2004, p. 287). In light of this, I present the following data and analysis to demonstrate how teacher evaluation policies, practices, and instruments have not only externally impacted teachers, but have also been taken up and embodied by teachers. In doing so, they have taken up a particular way of seeing themselves and their peers, and qualifying their effectiveness and worth, in relation to such policies, practices, and instruments. There are three prominent ways in which they have subjected themselves to such technologies of governance: 1) they define their value, worth, and quality in terms

of that which can be numericized; 2) they embrace and encourage hierarchical surveillance; and 3) they have discursively constructed a dichotomous view of what it means to be a good DMS teacher that is based on a willingness to comply with the latter two embodiments.

The teachers and evaluators who participated in these interviews were unique in many ways—they came from different backgrounds, years of experience, they taught different grade-levels and subjects, and they expressed different levels of comfort with and approval of TAP. Similarly, they each had unique experiences dealing with TAP, which was apparent in their different responses. However, along with such difference also came similarities. In keeping consistent with my theoretical framework, my responsibility as the analyst was not to assess each person's experience and determine who of the participants was closest to some truth; nor was it my intention to compare and contrast their stories in an effort to figure out who was more right or more wrong. For example, when asked to discuss the alignment of teachers' Student Growth Percentile scores with their observation scores, the principal said that she had seen many inconsistencies. However, the vice principal and master teacher said that the scores were consistent across the board. While a traditional policy analyst might take this discrepancy as a unit of focus, I was more concerned with how the participants framed their responses around such evaluative instruments and practices in order to understand how they made sense of teachers and teacher quality. Thus these findings are not meant to generalize to the rest of the school, district, or beyond.

I will first present the data in a case-by-case manner, demonstrating how each of the participants has conceptualized his/herself in terms of the evaluation techniques at

play. I begin with an analysis of each participant individually in order to keep the responses contextualized within each participant's experience. I will present each case by providing excerpts and brief analyses of such. My goal is to demonstrate how the teachers and evaluators have taken up a neoliberal discourse to define themselves, their teaching quality, and their colleagues (e.g., the evaluators qualifying the teachers). Then I synthesize the findings to make sense of the cases as a whole, at which point I draw conclusions about how Desert Middle School educators have discursively negotiated what it means to be a quality teacher. I also attempt to link their conceptualizations of themselves to the technologies of governance discussed in Chapter 4 and argue that the participants have created an environment where teachers conduct themselves in particular, desired ways, while simultaneously confining quality teaching to a narrow set of criteria. In an effort to protect anonymity, I have replaced all of the participant names with pseudonyms that were generated randomly with Scrivener writing software.

Case 1—Christina (Career Teacher)

At the time of the study, Christina was a first-year band teacher who previously taught as an adjunct professor before joining DMS. She also spent several years teaching private band lessons, which is how she discovered her preference for middle school students. She said that she plans to continue teaching, stating: "I feel like I finally have found what I'm supposed to be doing." During our first interview, Christina had only experienced one cycle of evaluation, but when I asked her about her earliest impressions, she stated: "My first impression was kind of grateful…I heard they were going to come in four times a year, I thought that was already something better [than what she had experienced with her student teaching]."

In describing her experiences with TAP, Christina expressed an appreciation for frequent observations (both announced and unannounced), as well as the way in which the TAP system brought a level of validity to her position as a band teacher. When discussing her preparation for her observations, she stated:

> I think right now it is time well spent because I'm still trying to figure out, I guess my place in the rubric, you know, and how I fit in as a band teacher, um, and I want to justify, I don't want people to think that I just stand on the podium and music comes out, you know, that would be wonderful if it did, but it doesn't, like I have to put in plans, and I'm interacting with the kids just as much as any other teacher, as any other good teacher should. So if I can say, like, yes, we're doing this this and this, it makes my job more valid, and it makes my job more professional, and hopefully less likely to cut if there ever was a question of being cut.

Of particular interest here is that, even though Christina expressed a sense of challenge to teaching music ("I [don't] just stand on a podium and music comes out"), she also communicated a need for rubrics to validate her teaching role. The rubrics allowed her to turn her process into a technical, numerical representation that was capable of being examined by an evaluator. Similarly, she used the rubric as a way to justify her position as being both worthy and professional in case of any budget cuts. Music, in this sense, was not important for music's sake, but for its capability of being professionalized through a meeting of standards and being numericized and evaluated. Christina furthered this justification when she discussed what she learned from her first observation to apply to her next one, stating:

> After the first one, I realized, um, I think what my areas that needed working on or

> whatever was thinking skills in the music room, and what I gathered from that conversation was I just didn't know how to justify what I was doing in the classroom. So we do all this igher level thinking in music, but the way that it's worded out like in the TAP rubric, it's kinda hard to talk your way around it, you know? So I went online, and I immediately got a spreadsheet for myself for the second one and was like there are the higher level thinking that they're doing, like they are reading text, and [inaubible], they're communicating with one another and it's not through voice, it's through music, and they're doing so much more than you can see, and, you know, I have to point it out more, and for me I take a lot of that for granted, so I've had to do more research of like what is actually going on.

Above, the focus (and time) was less on the actual music and more on how to translate her teaching of music into something visible and thus audit-able—something that could be seen, measured, and assessed. She also stated:

> I do think that if if we have a strong music program, that's going to bleed out to the rest of the school and help everybody. Either by bringing more kids that like band, you know, or having, I don't know I don't think band makes you smarter but I think it attracts smarter kids, so that could up that score.

Above, Christina defined the significance of music in terms of its market value. In her view, a better band program might attract smarter kids, which would be a marketable asset for the school; and it would create the potential to raise overall test scores.

Case 2—John (Career Teacher)

At the time of the study, John was a second-year teacher who got into the profession via Teach for America (TFA). Before teaching, he was an attorney in family

law, where he represented children in the court system. After too many divorce cases that "felt more like hurting than helping anyone," he decided to join TFA and become a teacher. Throughout the interview, I noticed how John's background colored his experiences as a teacher. When I asked him to talk about his general thoughts on TAP, he responded:

> I had heard about TAP and programs like that, like teacher, like performance-based pay initiatives for a long time, and it just seemed like um the next logical step in teacher accountability, so I mean I wasn't like shocked or surprised or anything.

In the above statement, John discussed teacher evaluation from a different vantage point than the other participants—one as an outsider looking in. His reference to the "the next logical step" demonstrated the widespread acceptance of viewing teachers as market entities in need of higher accountability and monetary incentives. However, when John talked about the student growth component of the evaluations, he explained that his former outsider perspective changed after seeing it in practice. He stated:

> The student growth I don't consider, like that's what I was all about when I walked in the first day. I was like, I need the numbers, like I'm going to, I have like I have a plan for these kids, I will judge my success based on the numbers at the end of the year. Then I got the numbers and then looked at the numbers for the past few years, and I did exactly what the teachers have been doing in my position for like the past 10 years, like it's not changed. It's shifted maybe a percent one way or the other, but really nothing's changed since I got there. And nothing's like it's going to be the same after I leave. So, it's about, so then for me it's about qualitative stuff. It's about what I can grow in as a teacher, how I can be more effective, because I mean, if you

> can't shift the numbers, which history has shown us that you can't, then what do I do then? And the answer is like be as effective as you can.

In the response above, John realized that though quantifying learning via standardized tests made logical sense to him, it was quite different in practice. Instead, he called for more focus on "qualitative" measures of teacher effectiveness; however, later in the interview I asked him about the fairness in the system, and his response suggested that while he might not have full faith in the system, the ways in which he thought about teaching and learning were still tied up in a market-based discourse:

> I think you just, you have to measure performance somehow, and that's a standardized test, and teachers' job performance has to somehow be tied to that. I think that it will lead to education reform in Arizona. I think that is the formula. Whether it's fair? Or totally accurate or perfect? I would say maybe not, but there's no other I mean (laughs) what other job are you not based on what you produce? There's no other job that you're not measured on that standard, and, like it may be unfair, it may be inaccurate at times, but that's all you have is what you produce, and you have to produce the best product that you can, which is who can perform on a test.

Above, John's reference to the "best product" demonstrated a common conception of the teaching-learning relationship as it had been defined by a market-based model of education. As mentioned in Chapter 4, teachers are producers of the product, and students are the consumers.

Case 3—Mary (Career Teacher)

At the time of the study, Mary was a fifth-year art teacher in her second year at

Desert. She received her certification traditionally, but before she began teaching, she was on the "Navajo reservation serving at like a private institution in a different capacity than a teacher." Before working at Desert, she was a teacher in Illinois where, in her words, "they have very strong unions, which are just evil." During the interview, Mary expressed her appreciation of the TAP system for its ability to get rid of bad teachers. She stated:

> One of my big pet peeves in IL was that they couldn't get rid of the bad teachers. Well, here, just that concept of you're constantly being evaluated and you're constantly evaluating yourself, like usually the bad teachers will leave because they don't want to be evaluated, so to me it just weeds them out right off the bat. And, they weed themselves out to another district that doesn't have as strict evaluations, and then to me that just makes that district worse, you know, cause they're just sitting there not trying to improve themselves.

Above, Mary referred to a common characteristic of audit culture—that of constant surveillance. In this case, she referenced a weeding out process, where teachers who were afraid of evaluation, or who lacked the desire to improve themselves, left on their own accord. Rose et al. (2006) argued that techniques, such as surveillance and audit, create "autonomized" and "responsibilized" subjects who perform in particular and desired ways. In Mary's remarks, she viewed the Desert staff as being responsibilized subjects because "when the TAP rubric started, they lost like 40% of their teachers, and it seems, they…it was the teachers that didn't want to deal with it."

Given that Mary was an art teacher, she was considered a "specials" teacher, meaning that 1) she was a Group B teacher, 2) her students got pulled from her class if

they needed extra AIMS preparation, and 3) her class was not factored into her students' overall grade point average. In discussing this, she talked about how she and her colleagues had to find ways of making their specials classes fit the rubric, stating:

> Well the thing we've come up with, um, just in our consultations with each other and in our own cluster meetings is like for music, when they're reading, they're reading music, you know they're reading the notes, it's a different kind of reading. Um, in art class, um, we might be reading visual cues, um, but also, like I had them read a story to go with what we were doing earlier this week, and then, um, since we were doing origami like you know there's written instruction, so there's the visual, there's the written, you know, like step-by-step understanding, so I still kinda get in some of that, but it is very project-oriented, project-based, and portfolio-based, which touches on in the rubric a little bit.

Of particular interest here is the way in which she discussed specials classes (i.e., art, music, band, computers, and physical education) in terms of how they related to the rubric, as if the justification of such subjects depended solely on their measurability (or evaluate-ability). Similarly, when I asked Mary how she felt about the idea that half of her evaluation was based on student test scores of which she had little to no control, instead of questioning the practice of using test scores, she told me about the district's discussion of a potential specials test, stating:

> This year would be just a year to see if we wanted to do it, like just to test it out, test out the test. And um, I think almost unanimously, we all were like, this is ridiculous because the kids are tested SO much, and now we're going to pull them to be tested for our areas too? And then like half of the test still said, well it's just

> observation based on your own personal observation, so then it's still not like concrete data necessarily. Like did they improve or did they not improve, or portfolio-based, like just seeing. And because of the fact that we don't have them for a full year, we only have them for a trimester, they rotate, so it's like, AND they mix the seventh and eighth graders. So then it's like you can't even do a seventh grade curriculum and then an eighth grade curriculum, it's just so (*) once again, the fact that they have the arts and the electives here is awesome. How it's done is still, it's just like you're a filler kind of thing.

In the excerpt above Mary disapproved of testing not because art was immeasurable, but rather because teacher observation of quality art was not sufficient enough. She considered art and the other electives as "filler[s]," reasoning that personal teacher observation is not "concrete data" and that "just seeing" does not really say anything about how well the students have done. This is consistent with a neoliberal discourse, usually related to terms like rigor and accountability, which have been commonly used to justify over-testing and the elimination of classes like the arts.

Case 4—Jennifer (Career Teacher)

Jennifer was a 20-year veteran teacher, 17 of which she had spent at Desert. During the study, she taught 8th grade English, which she had been doing for many years. As of the last interview, she was not sure that she would stay with teaching, stating, "the current system sucks the fun out of teaching." Unlike most of the other participants, Jennifer did not talk about the strengths of the school and the administrators in terms of TAP, but she talked about their strengths in *spite* of TAP. From the beginning of the first interview, she made it clear that she was not a "TAP-y" teacher and that she never planned to become

one. For that reason, she questioned whether she wanted to remain in the profession. Though she never spoke of her teaching or of herself in terms of TAP per se, she did qualify herself and teacher professionalism in terms of market-value terms, which I will discuss through the following analysis.

Of all of the participants, Jennifer demonstrated the most resistance to TAP, regarding the rubric as too restricting and too specific. She had been vocal about her resistance, stating, "Well, you know, I have never, ever been quiet about the fact that I think that this sucks the fun right out of it." However, within a governmentality framework, resisting to participate as a powerless agent does not mean that the subject is in some way free from power itself. Rather, the idea of a "resistor" is unthinkable without the something that can be resisted, just as all identities are made up (Hacking, 1999). As such, the "resistant" subject is constituted by the very policy, practice, etc. being resisted. In other words, there is no set structure of power relations that the subject can get outside of and act against.

Interestingly enough, even though Jennifer disapproved of TAP specifically, she had still embodied a market-based discourse, defining teaching and qualifying herself as a teacher in terms of her market value. In the following excerpt she talked about herself as a teacher:

> I pretend like the rubric matters a little bit, while, when I know I'm being evaluated. But the rest of the time, I teach. And I tell you what, the reason I'm here after 17 years is because I'm good at what I do. And what I do doesn't look like that, but I still get the highest, um, growth scores because I know how to teach kids, and it's because of how I teach. And it's not like a quant[itative], it's not something I can

> put in a bucket and give you some of, it's just this is what I do. I know how to teach that kid because when I talk to that kid I use this tone and I use this language, and I, you know. And then I get the exact same results from this totally different kind of kid by using a totally different set of everything. And this is way too cookie cutter. Sure I'm supposed to do differentiation between low, medium, and high, but that's not, it doesn't give me the flexibility to really address all 36 levels of kids in every class because there's no two kids that are on the same level. They're all in a different place at a different time, and if I'm trying to follow a procedure they way that I spend way too much trying to follow the procedure, and I don't pay any attention to what kids actually need in order to learn how to do the thing.

In the above excerpt, Jennifer explained that she did not like following the rubric because it did not help her address the needs of all of her students. Rather ironically, however, she still defined the needs of her students in terms their performance on standardized achievement tests, thus reducing her own effectiveness to that of a test number. In another part of the interview she justified her value to the school, stating:

> Yeah, are you going to FIRE me because I'm not going to dot my I's and cross my T's the way you want me to? Or are you going to keep me because I can teach the kids that nobody else wants to teach how to do stuff that they're not even willing to even attempt for anybody else.

Similarly, Jennifer was not resistant to TAP-like instruments and practices, such as observations, rubrics, and VAMs, just TAP itself. She stated:

> I'm not opposed to it [VAM] because I think, I mean, that's what they pay me to do is to grow kids. That's my job. I think they should shut up and leave me alone and

> let me do it in the best way I know how and I'll GET your numbers for you, and they'll do what they're supposed to do if you would just shut up and leave me alone and let me do this (laughs)…And I'm ALL FOR observations.

In speaking about being observed, she preferred the unannounced visits because only then were the observers able to get a "real picture of how someone teaches." She stated that she was a "little more I-dotting during the weeks she knows she will be observed."

Interestingly enough, when discussing her last observation/evaluation of the year, Jennifer, while remaining on the "people-who-need-to-get-their-scores-up-or-there's-going-to-be-a-problem list because [she does not] care enough," she began to equate her teaching quality in terms of TAP, stating:

> My lesson was much better and had much better numbers then they have in awhile because apparently I understand some things that I didn't, I wasn't, I don't, you know the way the stuff is worded, I don't, why did they have to rename everything (laughs)? And then it throws me for a loop, and I'm like well I'm trying to do what these words are telling me to do, but that's not apparently what they really mean.

Using the TAP example of "closure," Jennifer went on to explain that her bad scores in prior observations were not only a product of her resistance, but might be a matter of her not understanding the TAP rubric. However, she also stated that even though the rubric expectations state one thing, she thought it "takes some of the craft of the art and craft part of it out of it, and makes it a very technical and scientific thing, and teaching is not." So even though she referred to her "better numbers" as evidence of a better lesson, she still held that the TAP rubric did not represent quality teaching to her.

Though more resistant than the other participants, Jennifer was not free from a

market-driven discourse who defined quality teaching in market-value terms. She equated student learning to evaluate-able outcomes, stating: “I'm not opposed to high-stakes testing and using that as part of a teacher evaluation because it doesn't matter, I don't think it matters how they get the knowledge, just get the knowledge.” This case exemplified why the traditional concept of resistance is shown to be of less concern because though Jennifer might have resisted her confinement to the specific TAP expectations, she was not free from a market-based discourse in the way she thought about herself as a teacher and qualified her teaching abilities in terms of a market value.

Case 5—Sarah (Career Teacher)

At the time of the study, Sarah was a first-year, seventh-grade science teacher. She was with Teach for America (TFA) and had completed her undergraduate work at Harvard University. She stated that when she began her TFA position she had planned to teach for five to six years, but that she soon realized that “it's hard to change things being a teacher,” so she planned to “go the grad school route after [her] two years are up.” She stated that she wanted to be a teacher so that all students could have the chance to go to Harvard: “you have to prove to me that you can do work, and you have to do work, you're not like going to magically go there. But, nobody told me that, so like people need to let these kids know that.”

Overall, Sarah expressed support for numericizing (Rose, 1999) teacher quality, stating: “I think that assigning it numbers is the best way that we have to do things right now.” However, she also called the system “reductivist [sic],” saying it “doesn't tell the whole story.” This dilemma was further complicated by Sarah’s expressed conflict between what she considered the “best type of instruction” versus what was “best for

[her] students." She said:

> I have to kind of conform to the rubric and do it their way, which is not necessarily the best way for my students. Do the TAP categories, totally make sense? Yes. And ideally, is that the BEST type of instruction? I think so. Like I really do like constructivist learning. BUT, it's not perfect for every class. My third period? 34 students, seven of which have IEPs, I'm going to do a lot more direct instruction, they're going to have less time to discuss because there's ALSO seven kids with behavioral plans in that class. And you can't, I can't give them that structured discussion time, or as much. So it's kind of like, once you know your kids, make it as TAP-y as TAP-y possible, but like, TAP doesn't necessarily translate to wonderful instruction for each child and what each child needs.

This conflict is interesting in that Sarah equated quality teaching, or the "best type of instruction," with an idealistic way of teaching instead of what might be best for students. As such, she separated best instruction from needs of students. This paradox begs the question: best for whom, or for what? Similarly, when asked to describe an ideal evaluation system, Sarah said:

> I think it needs to be standardized, and I think it needs to be measurable and you need to get results that are accurate or precise or whatever, and that can be compared across schools. However, at the same time, I think maybe in order to do that … it would need to have fewer indicators on it. And it would be more of like students got it, students didn't get it, here's evidence that shows us the things that you did that you know your kids and gave them what they needed. BUT THAT, TAP, like I said, it's already subjective, and you, that becomes even more

subjective. So how do you get quality data from that? I don't think you really can. Above, she critiqued TAP in particular, but she called for an even further numericized system that would eliminate all human judgment. Again, quality was reduced to something measurable and evaluate-able, making teachers subject to audit (Davies & Bansel, 2010). Sarah, though concerned with the TAP system specifically due to added stress and its misalignment with students' needs, expressed her support for a "standardized and measurable" system.

Case 6 — Nicole (Career Teacher)

During the time of the study, Nicole was an 8th grade mathematics teacher in her sixth year of teaching and had worked at Desert since the beginning of her career. She spent the year prior to the study working as a mentor teacher who was responsible for teaching her own classes while also coaching and evaluating her peers. She decided to leave that position and return to the classroom as a full-time career teacher. She stated reasons such as pay, time, and stress for her decision to leave the mentor position. Of all of the participants, Nicole expressed the most support for the TAP system, specifically the rubric, stating:

> [TAP] made me more aware of things that I didn't think about before….Since year one 'you're doing fine everything's fine.' And I always wanted more, so I think the only place I found it was from the rubric. So I took the rubric like, like it was like gold or something. So, I, you know, had to basically teach myself to go through these things and so I became a much better teacher because of the rubric….[The] more knowledgeable you are of the rubric, the higher you're going to score.

Above, Nicole expressed her liking of the rubric because it provided her with ways of

improving, however, she also equated such improvement to higher scores. In this example, quality was expressed in terms of rubric-based scores, not necessarily better teaching. She then stated that she also saw improvements in student learning: "As long as I was focused on all these things, my lesson was coming together and the students were learning at a much better rate." But she did not explicitly define student learning beyond "their involvement in [her] lesson."

Similarly, Nicole valued student growth scores over the rubric-based scores, stating: "I think [student growth scores are] very important. I honestly think it was, like, it was more, the student growth. I mean I think that says a lot. It says a ton. You know, you can go and you can put on a show for someone easily." In this excerpt, true quality teaching was measured by the student growth, while observation scores were manipulatable by the teacher's performance.

Case 7—Melissa (Career Teacher)

At the time of the study, Melissa was in her first year of teaching. She was a special education (SPED) mathematics teacher who got into the profession via TFA. She stated that she intended to continue her career as a teacher well beyond the two-year TFA requirement. In relation to the other participants, Melissa's interview was the least common of the group. To start, her situation at Desert was unique in that, at the time of the study, she was working towards her certification as not only a mathematics teacher, but also as a SPED teacher. With this being the case, she was subjected to three different types of observation and evaluation—TAP, TFA, and The New Teacher Project (TNTP, through which she was working towards her certification). As such, she was responsible for teaching to three different rubrics at any given time. She discussed this challenge,

stating that the rubrics were philosophically and practically different, but that her main priority was getting her teaching certification, so she put most of her energy into the TNTP rubric. However, of all the participants, Melissa appeared to give the least weight to the rubrics (and other evaluative practices and instruments) when she discussed herself as a teacher and her teaching practices. In talking about a recent lesson, she explained:

> Like yesterday we just went outside and walked around the area in the back and just found we're working on relationships with angles so we went around and we identified right angles, obtuse angles, acute angles, and then they kind of like organically like an inquiry lesson I guess were like I know this is like 180°and this is a crack in the sidewalk so every time there's a right angle there's 90° and 90°that's a line. And they're like 'oh my gosh, but this is this is an acute angle every time it's not a right angle, that means this one has to be the big one, which is obtuse.' Yeah the students the day before who told me that like the line is an angle. We've come so far. It was good….It's so funny that I'm telling you this because I have no idea if that's okay, do know what I mean? Like I have no idea. Like I'm way in the back of the school and nobody ever comes to my class. I don't even know if this is okay, but I feel like we're learning for the first time ever, so, like all of us. So we're just going to keep doing it. According to the rubric I have no idea how stations are evaluated we don't really talk about that in our meetings so much.…I don't know how it's going to work on the rubric though because it's not traditional. At the beginning of the year somebody told me that the resource class should be taught just like the gen ed classes are. In my mind and like my naïve like intro, I do, we do, you do, exit ticket. And the kids were like (gestures that it went

over the students' heads)….For this way is what's working but how that fits into the tap rubric I really don't know. I'm really just trying to do what's best for my students. And I'm excited.

In the excerpt above, Melissa defined learning in a different way than her colleagues. She did not qualify it in terms of a standard or a test score. Instead, she equated it, simply, with the students understanding the concept and applying it to something in real life. Her only way of knowing that they learned the concept was by her own observation and judgment, which she did not present as unworthy. Similarly, when referencing the rubric, she downplayed its priority in her decision-making, even going as far as to admit that, regardless of what the rubric stated, they are "just going to keep doing it."

Case 8—Robert (Master Teacher)

At the time of the study, Robert was a master teacher at Desert Middle School. He was responsible for teaching one 8th grade literacy honors class and for coaching and evaluating the language arts and social studies teachers. He also was responsible for planning and facilitating the weekly cluster meetings for the teachers whom he was responsible for evaluating. During the time of the study, he was in his eighth year of teaching, all of which he had spent at Desert. This was his second year serving as master teacher. Though he expressed dissatisfaction for the money component of being a master teacher because "it's only like 100 bucks a week that you make being a master teacher," he also expressed enthusiasm for TAP and his role in the system.

Robert's explanation of TAP and its operation at DMS was consistent with a neoliberal discourse, specifically in terms of responsibilization and audit. When I asked him about the relationship between the teachers and the evaluators, he brought up that

almost half of the teachers left after the first year, stating:

> The first year was rough. I think that year we turned over half the staff. A lot of the old teachers who didn't want the accountability and weren't, it was self-conscious, which is fine, I mean, but you know, TAP was kinda the first, we got a bad wrap for it, but right after that came student growth models, so either way, if you're not going to be found out in your observations fine, but observations with those numbers of kids growing is going to be telling, so, so either way you have it.

There were two keys points of interest in the above excerpt. First, he associated the high turnover rate with accountability, making the assumption that teachers who left were afraid of accountability. McWilliam and Jones (2005) argued that "being within sight is the sine qua non of the professional," (p. 117). Thus those thought to be unwilling to comply with such exposure were cast as unprofessional, or, in this case, afraid of being "found out." Similarly, in his use of the phrase "either way…found out," he implied that a teacher's worth is based on either rubric scores or student growth scores; thus limiting the way in which one can be a quality teacher.

Robert expressed a similar position when he discussed his own position within the TAP system, stating:

> My first year I was, I loved the looseness and the independence and the no one really checking on me. At the same time, when you create this atmosphere of like this, I don't know how to say this, I'm just going to say it, you create this give-a-shit attitude. Like if I'm doing a great job, great, no one knows about it, no one gives a shit, no one's going to call me on it. If I'm doing a bad job, great, no one knows it, no one's going to call me on it, no one gives a shit. And maybe that's not the

attitude, but that's sometimes portrayed when there's no accountability and there's nothing in place.

In the above excerpt, surveillance was a means of getting the teachers to do their jobs, otherwise they would not have a "give-a-shit attitude." This statement implied that teachers do not work hard for the sake of their students, but rather for the sake of the teachers' susceptibility to exposure. Similarly, when referring to teachers, he used the phrase "they were only a 2.5 teacher," which directly defined the teacher in terms of a number and placed the value of the teacher on that number. This reinforced the idea that teacher quality was reducible to simply a number.

Case 9—Heather (Master Teacher)

At the time of the study, Heather was in her third year as the mathematics master teacher at Desert. Before entering the master position, she served three years as a career teacher and one year as a mentor teacher, all at Desert as well. She got into teaching via TFA. Similar to other teachers, Heather discussed the high turnover rate at the beginning of TAP implementation, but she called it "a healthy parting that actually made our school stronger. It was really difficult at the time, but in the long run it's been beneficial." When I asked for clarification, she stated:

I think the staff that stayed really viewed the rubric and viewed the observations as like a way to get their instruction stronger, um and just seeing the transition in instruction on campus has been super powerful to see. Um teachers really just recommit and refocus what they're looking for in a lesson. Um, and some of the teachers we lost were not ready for that. So I think ultimately it was better for kids, um, with who stayed, and ultimately who replaced a lot of them.

This was another example of how the teachers were expected to be open to exposure and how, as a campus, they had come to see this practice as necessary to the extent that teachers who left were assumed to "not be ready" to improve. Heather also discussed another form of surveillance that was related to making decisions and intentions viewable, measurable, and audit-able. She called this being "consciously competent" and it explained it as such:

> One of the things that we learned with TAP, which was interesting, was the idea of like consciously competent, so someone who knows what they're doing and they're conscious of the decisions they're making. …It's very clear in the lesson. They'll say, we're going to work in these groups today because I want you to do blah blah blah, so the kids are aware of it, the observer is aware of it, and it's clear that the teacher is, so. I think there's a difference, and so, the evidence in the classroom really shows it, and then in the lesson plan too, you'll see what they're taking note of versus what they're recording just for the sake of recording and what they're recording because they want you to know the decisions they've made and why they've made them.

This was different than explicit observation in that an evaluator did not have to walk into the teacher's room. But the teacher was expected to make his/her thinking, or intentions, explicit and visible, thereby creating a different kind of exposure. Particularly, the comment "they want you to know," implied that the teacher was not necessarily writing out his/her intentions for the sake of teaching or learning, but, rather, to show the evaluator what he/she had done and why. Heather also discussed the expectation of constant, explicit observation and the difference between the new teachers and veterans:

> A lot of our new teachers are either coming out of iTeach with ASU or TFA, and so they have been schooled in a system of constant observations. Um, like seventh, like my people that I'm observing next cycle, they're like, oh come in whenever you want, I'm so used to people coming in, like whatever. So their culture of like opening their classrooms to observers is way different than some of our veteran teachers who have been here awhile and the observation culture then was maybe once a year maybe once every other year.

This reference to external institutions as potential reasons for teachers being open and comfortable with constant observation reinforced the idea that desired ways of being a particular teacher was not unique to Desert, but rather a manifestation of a neoliberal discourse that has shaped every aspect of the education system.

Case 10—Lisa (Vice Principal)

At the time of the study, Lisa was the Vice Principal who began her career at Desert as a teacher before moving into her leadership position. In terms of TAP, she was responsible for the exploratory teachers (i.e., band, music, art, computers, and physical education), which she attributed to her prior experience as a physical education (PE) teacher. In addition, she also observed and evaluated almost all teachers on campus. During the interview, Lisa spoke highly of TAP, especially for its impact on the school culture, stating:

> And so I think that the teachers just see that as okay, we're all in this together, we're just trying to get to the ultimate goal, and so, if there are people in my classroom watching and giving feedback, they're very open to it. And I don't know if maybe it's the relationship we HAVE with them, like our master teachers are WONderful,

and our mentor teachers are awesome, like we have some of the BEST.

Above, Lisa provided a rationality for why teachers should be open to having observers and evaluators looking over them at any time. This rationality, which was also expressed by others at the school, created a dichotomous view of teachers—you are either with us or against us. Davies and Bansel (2010) found a similar phenomenon in their study of governmentality and academic work, writing:

> The self-interest of the academic is re-constituted in terms of the interest of the university, and the self-interest of the university translates back into the interest of the academic. These acts of translation install the interests of the institution at the heart of these transactions such that those who do not comply put the institution itself at risk. Conformity thus acquires a moral imperative larger than one's personal survival as an ethical being, (p. 9).

I also asked Lisa to talk about the exploratory teachers and how she saw TAP work in their classes. She discussed how the PE teacher has done an exceptional job at making TAP work in PE class:

> I'll go to our PE teacher, our girls PE teacher, she's been here four years, and every year she's gotten better, like her objectives are posted, and her assessments, they have rubrics tied to them, and she has kids assess one another on different skills, and clipboards out and they're assessing one another, and it's just the idea of okay, what does this look like, what can the possibilities be?

In the above example, Lisa's phrase of "she's gotten better," was directly linked with the teacher's ability to conform to the TAP rubric expectations. By doing this, she minimized the teacher's quality as it related to other goals that might be involved in teaching PE. I

asked Lisa if the PE teacher was a better teacher because of this, and her response was:

> Yes, I think, it makes what she's doing more focused on like the outcome, so instead before they might have done all of these things, but hadn't given each other feedback, or hadn't like evaluated one another, so I think it just pushes it to the next level of her understanding of what they can do and ultimately demonstrating what she wants them to do, so I think it does make a difference.

Again, she linked teacher quality with TAP expectations rather than anything regarding students' "physical education."

Case 11—Becky (Principal)

Becky was the principal at Desert Middle School. In order to contextualize her interview and illustrate her presence on campus, I want to first tell the story of my first meeting with Becky. The first time I met her was on a morning before school. Since she normally spent her mornings monitoring the campus, she invited me to walk with her. She grabbed a whole, raw carrot (saying, "I'm trying to get in more vegetables"), and we made our way to the courtyard where the kids played and danced to loud music. She explained to me that she wanted students to *want* to be at school and that having a little fun helped them focus by the time the first classes started. Her interaction with the students made it clear about the type of culture she had worked to create—one of community and respect. This sentiment was reinforced in several of the interviews I had with teachers.

By the time I officially interviewed Becky for the study, it was at the end of Desert's fourth year with TAP. Having been at the school for more than a decade prior to TAP, I asked her to describe some of the changes since its implementation. One of her

responses dealt with measuring teacher quality before and after TAP:

> I think uh, um some of the frustration of they were, they were good before, but we had never really been able to measure the effectiveness of their teaching except once a year when we got the AIMS scores. And then you never really had the chance to look back and go, okay if we want to be better in the classroom, what should we work on? And the rubric actually gives us specific things to work on.

The above statement reinforced the idea that teacher effectiveness was defined in terms of test scores. The rubric here was positioned as a tool to improve student achievement on AIMS because it allowed teachers to know how to better prepare the students before the test. In talking about student growth specifically, Becky stated:

> I think that's the most important thing that the beginning of NCLB is say, you know we actually need to figure out if our kids are improving or moving forward, and so it took away, starting to measure whatever it was they wanted to measure um, I mean it was a nice philosophy, but it bombed, but it has started us down, it started the district down the right path when we started to be able to measure growth and seeking it at a high level.

Above, she further defined student achievement (and thus teacher quality) in terms of scores. The statement "we actually need to figure out if our kids are improving," reduced learning to a very narrow construct that could be numericized and evaluated.

Becky also talked about teachers who were "perfectionists" and, regardless of their evaluation scores, would continue to self-improve:

> And regardless of what an observer says really it kind of doesn't matter, um, if you have a bunch of people who have high expectations for themselves and they're

> perfectionists, is not acceptable. It, they could get, they could get all 5s and one 3 and be devastated, okay? Or one, (pretends to hyperventilate), all fives and one four and it's just not good enough. And even if you gave somebody all fives, and we have some teachers that are close to that, they're phenomenal because of the um turning over of their classroom to the kids and basically they're just facilitators, and for them to get to that point, they're just that good. But even if you gave them all fives that would not be good enough because they still in their hearts know of something that they could have done better. But that's the type of people that you want.

Power here was acting in a way that did not require a figure of authority physically mandating the action of someone. Instead, by way of normalizing, or standardizing, practices, teachers were encouraged (and desired) to practice self-regulation. "Self-regulation occurs by virtue of a norming process whereby the power of societal norms is internalized by educational system participants," (Foucault 1984, as cited in Graham & Neu, 2007, p. 312). In the following sections I will provide the cross-sectional analysis, whereby I will discuss the teachers as a collective and map the teachers' responses onto the evaluation technologies of governance at play.

The Audit-able Teacher

> Audit technologies are a means of governing subjects; of making them more governable by constituting them as the sorts of subjects demanded by the programmatic ambitions of government. In being taken up as one's own ambitions, the ambitions of government become technology of the self, (Davies & Bansel, 2010, p. 9).

The teachers and evaluators at Desert Middle School had taken up a discourse that encouraged and evoked a sense of 'entrepreneurial actors' (Brown, 2003, p. 38), or teachers who were valued in terms of their numericization, surveillance, and willingness to subject themselves to such practices. As such, the teachers governed themselves, or behaved in desired ways, that constructed "the type of people that you want," (Becky, principal).

According to the participants, only teachers who want to improve have stayed at DMS—other teachers have left in fear of being "found out," (Robert, master teacher). This socialization of teachers re-constitutes teachers' ethos in terms of their willingness to comply with the technologies of governance. The attitude of the teachers was re-constituted in terms of their interest of the overall school. Similarly, teachers were viewed as valuable only if they were willing to be subjected to audit, all in the name of school. For example, Lisa, the VP, stated: "we're all in this together, we're just trying to get to the ultimate goal, and so, if there are people in my classroom watching and giving feedback, they're very open to it." Here, surveillance is rationalized for the sake of the "ultimate goal." In other words, teachers unwilling to subject themselves to surveillance were positioned as teachers who do not care about the school.

Teachers were viewed dichotomously as either acceptable or unacceptable. The "good teachers" (Fenwick, 2003) wanted to be observed, wanted feedback, and were willing to sacrifice certain individual freedoms for the greater good. Bad teachers were painted as scared of surveillance, for they either did not want to improve, or were bad teachers and did not want to be caught. As such, the teachers saw their peers and themselves in these ways, which encouraged them to adjust their own behaviors as to not

be one of the unacceptable ones.

At DMS, there were two technologies (i.e., governing techniques) that were involved in making up (Hacking, 1999) the audit-able teacher (see Figure 2). First, teachers had to be numericized (Rose, 1999). Their practice, quality, and worth had to be quantified into something that was measurable and evaluate-able. Second, teachers had to subject themselves to surveillance by way of explicit observation and other forms of implicit examination (e.g., the submission of lesson plans, pre- and post-conferences, self-reflection forms, etc.). These technologies had their greatest effect on reported teacher behavior in the way judgments about teachers could and have been standardized, or what Foucault called "normalizing judgments" (Foucault, 1975). Given that the teachers had been provided and trained on a standard way of behaving (via rubrics, test scores, etc.), there was a way to judge their behaviors as being either normal or abnormal. Again, this also provided the teachers with a standard of which to compare themselves against and modify their behaviors as needed.

Audit by Numericization

> Whenever measurements are made, and results are aggregated, compared, and publicized, the result is the same: those who are the subjects of these measurements are revealed in their attributes, and they, therefore, adjust their behaviour towards the group norm, (Graham & Neu, 2004, p. 311).

With the use of evaluative instruments (i.e., SGP models and rubrics), DMS teachers and their practices have been quantified and made measurable. The process of measuring something is always subjected to and limited by (1) those who choose to measure it, (2) what, specifically, they choose to measure, and (3) how they choose to

measure (Rose, 1999). This is always at the expense of that which is not or cannot be measured. When teachers are reduced to such numbers, they begin to think of themselves in such a way, as evidenced in the teachers' responses.

In this way, quality has been constructed to be something that is not only evaluate-able, but also, once accepted and normalized, has become what Foucault (1980) called a "regime of truth," (p. 131) and not a truth itself. In other words, even though the numbers produced by these instruments are subject to the *what is* and *how* they are measured, the outcomes begin to define that which is measured. As such, the idea of teacher or teacher quality is made up in terms of such numbers. This numericization produces three possibilities: 1) the possibility to objectify and thus compare teachers, 2) the possibility to define teachers in terms of their market value (e.g., the rationale behind value-added models), and 3) the possibility for teachers to be subject to discipline, such as the rationale behind "measure and punish" (Amrein-Beardsley, 2014). Simultaneously, other possibilities of being a teacher might be eliminated.

For example, when the master teacher referred to someone as "only a 2.5 teacher," he explicitly defined that teacher in terms of a number, thus eliminating all other characteristics that cannot be measured. Also to note here, even scholars who do not necessarily focus on the discursive issues with numericization and more concerned with how statistics can help evaluate teachers, argue that value-added scores, such as the 2.5 in this case, can be quite arbitrary (Berliner, 2013; see Holloway-Libell & Collins, 2014). However, this evaluator, as well as several of the teachers with whom I spoke, appeared to have accepted this characterization as true and definitive. Similarly, in terms of market value, the numericization of practice creates a rationality for professionalism as well as

validity in the case that cuts are necessary, as was articulated by Christina, the band teacher: “So if I can say…we're doing this this and this, it makes my job more valid, and it makes my job more professional, and hopefully less likely to [get] cut if there ever was a question of being cut.”

In terms of numericization producing the possibility for discipline, this was evident in the interviews as well. Discipline, here, has multiple meanings—to punish, or to behave in particular ways (i.e., self-discipline). Mary, for example, described a teacher who resigned midway through the prior year:

> There was a TFA teacher here last year that, granted I don't know her whole story, and she was a real sweetheart, she wasn't bad or anything, but I just don't think she could hack it as a teacher. I accidentally walked in when she was getting her post evaluation with the principal and she was kinda crying and so I felt, like even talking about, I just saw her emotion…and then she decided to leave halfway through the school year.

Mary’s use of the phrase “don’t think she could hack it as a teacher,” rationalized the use of numbers to discipline teachers. As a result, teachers act in certain ways, or discipline themselves in certain practices, to avoid external discipline. For example, when Jennifer was discussing her third cycle of observations, she said, “My lesson was much better and had much better numbers than they have in awhile because apparently I understand some things that I didn’t [before].” Jennifer, who was the most vocally opposed to TAP, explained that she changed her practice, and received higher scores. She acted in the (TAP) desired way, despite repeated personal criticisms of such behaviors. Had she not, though, she likely would have either lost potential bonus money, or possibly even lost her

job.

Audit by Hierarchical Surveillance

> Technologies of audit and surveillance, of self-audit and self-survellience, are not simply discourses of responsibility and accountability but technologies for the production of responsiblized and accountable subjects. We think, then, of auditing as not primarily concerned with organizing and managing finances and outputs, institutions and workers, but as producing specific sorts of worker subjects (Davies & Bansel, 2010, p. 9).

I am defining surveillance as a means of explicit and/or implicit forms of observation and inspection. First, explicit forms relate to practices such as formal classroom observations and informal walkthrough observations. Formal observations are structured, planned, and consistent. At DMS, the teachers were formally observed four times a year—two of which were announced (i.e., planned with the teacher), and two of the observations were surprise visits but were confined to specific times of the year. Each teacher was observed by the principal, vice principal, master teacher, and mentor teacher. Informal classroom observations were random and lasted approximately five minutes in length. Teachers were unaware of when someone might enter their classroom, and it could be conducted by any of the leadership team members. In reference to observations, Mary stated the following:

> If they could all be unannounced, then I think they would be more realistic about where teachers are at. For me, I would prefer that because, I mean, if we actually taught on a daily basis, like the rubric says, I think we'd all be better teachers, quite frankly, so um, but unfortunately I've heard of the dog and pony show, and why

doesn't you know your scores match your evaluation scores, you know and I know with myself, with other teachers, like when we know an evaluation is coming, we go above and beyond at that time, and then when the unannounced are coming, they go above and beyond for a whole a month at least.

In the example above, Mary suggested that teachers did not do their jobs well or ethically without the fear of someone catching them at any given time. Other participants used another rationality for observation in the name of improving. For example, when Lisa (the VP) said that "teachers just see that…we're all in this together, we're just trying to get to the ultimate goal, and so, if there are people in my classroom watching and giving feedback, they're very open to it," and when John compared the type of feedback he got as a lawyer and the type he got as a teacher: "[they were] completely different. I did have, like bosses used to yell at me when I was an attorney, scream at me for doing something wrong, and I'd prefer that than [the feedback I get as a teacher], yeah, it's miserable. But it's like a necessary, a very necessary." This rationalized the practice of observation so as teachers are made to feel that it was not only helpful, but also "necessary." This was directly related to the way in which the participants talked about those who left because they were scared of being observed or afraid of being caught.

Another method of surveillance at DMS was more implicit and required teachers to subject their lesson plans for examination. For formal observations, teachers had to submit their lesson plans to their evaluators. They also had to collect other artifacts to demonstrate competencies such as "thinking" as per the TAP rubric. Additionally, teachers were required to participate in pre- and post-conferences to further explicate (or make visible) the decisions behind their teaching practices. Nicole, who was a mentor

teacher but then went back to the classroom as a career teacher, said the following:

> Being in the mentor role, that was my way of growing. I've kinda always been responsible for my growth, um, but then, it kinda turned on me. It's kinda like when I would get evaluated by my TAP team, it's like suddenly the pre-conference doesn't matter, the post-conference, it just kind of goes away.

Nicole looked at the conferences as an opportunity to grow, while also suggesting that without it, she could not improve as a teacher because nobody was willing to critique her practice (in a previous statement she said that most evaluators told her she was doing fine). Christina talked about a method of making visible her practice in the following statement:

> I immediately got a spreadsheet for myself for the second one [observation] and was like there are the higher level thinking that they're doing, like they are reading text, and they're communicating with one another and it's not through voice, it's through music, and they're doing so much more than you can see, and, you know, I have to point it out more, and for me I take a lot of that for granted, so I've had to do more research of like what is actually going on.

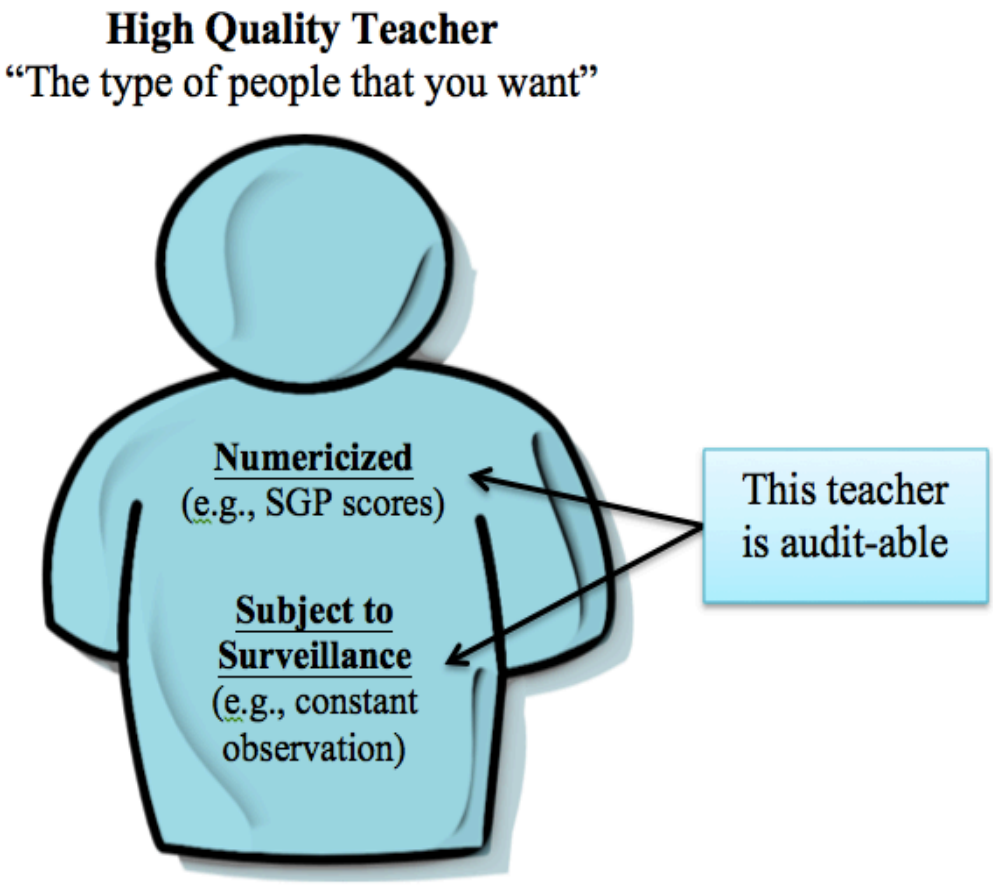
As the band teacher, she had to take extra time to make her practice visible because, to her, critical thinking was something different than what the traditional expectations might look like. As such, she had to prove that what she was doing was worthy of praise, begging the question: does this make her better at teaching music? The point here is that only through making the invisible (e.g., her intentions, decisions, etc.) visible, could her worth as a teacher be validated (by examination and evaluation).

The Un/Acceptable Teacher

The teachers and evaluators at Desert Middle School discursively constructed a dichotomous view of teachers, and in doing so made up two types of teachers—1) the acceptable teacher who was on board with TAP, and 2) the poor quality teacher who was afraid of TAP or did not understand its purpose. The high quality teacher (see Figure 2) was a person who was audit-able, which was made possible with two technologies—numericization and surveillance. Teachers who had subjected themselves to both numericization and surveillance were regarded as acceptable teachers. As teachers adopted this discourse and thought of themselves as numbers and subjects of observation and normalized judgments (Foucault, 1977), they modified their behaviors to be a part of the "normal" group. As such, they performed in specific and desired ways, so as not to be seen as "abnormal." Simultaneously, by doing this, they potentially marginalize, or even eliminate, other possibilities of being a quality teacher. "Ethics, here, was understood in terms of technologies of the self—ways in which human beings come to understand and act upon themselves within certain regimes of authority and knowledge, and by means of certain techniques directed to self-improvement," (Rose et al., 2006, p. 90).

Figure 2

The Acceptable (or Quality) Teacher at Desert Middle School



Simultaneously, teachers avoided the undesirable characteristics of the poor quality teacher. However, as Foucault calls this a "regime of truth," these "types of teachers" do not represent true, real teachers necessarily. For example, the characterization of teachers who left Desert the first year of TAP were described as afraid of accountability, or scared of being exposed through evaluation. This narrow view of possibilities as to why one might choose to not participate in TAP only reinforced the necessity for teachers to

behave in the normal, or acceptable ways. Out of not wanting to appear afraid, teachers had welcomed observers into their classrooms and minds willingly.

Table 3 shows the binary characteristics of the acceptable and unacceptable teachers as constructed by the teachers at DMS—the characteristics of each were taken from the participant interview data. This binary helped to construct self-governed teachers who behaved in desired ways through self-reflection and self-discipline.

Table 3

Constructed Versions of the Acceptable and Unacceptable Teacher at DMS

Acceptable Teacher	**Unacceptable Teacher**
Wants to improve	Does not want to improve
Is eager to be observed and evaluated	Does not want to be "found out"
Wants feedback	Feels "threatened" by feedback
Puts the interest of the school above oneself	Only cares about self
Is competitive with self and others	Is not motivated by competition
Believes in TAP	Does not understand TAP
Makes TAP work, regardless of class	Does not understand the rubric
Makes TAP work, regardless of students	Does not hold high expectations for all students
Is proud to demonstrate lesson planning processes (e.g., the "consciously competent teacher"—Heather, master teacher) (technology of the self)	Is a "shower planner" (i.e., plans the lessons the morning of class) and does not want people to know (Robert, master teacher)
Self-reflects and self-regulates	Is lost without guidance
Has high expectations for self	Has low expectations for self
Always striving to be better	Is satisfied with performance

It is important for me to note here, however, that this table is not to say that the characteristics on the left are not good qualities, or that the characteristics on the right are not bad qualities—in fact, I am not attempting to qualify any of these characteristics in any way. I am, however, arguing that these characteristics have been used to "make up" certain types of teachers and to elicit certain types of behaviors from teachers. For example, if teachers who have resisted TAP are labeled as "scared" of accountability, or in fear of being caught for bad practices, then in an effort to not be labeled as such, teachers might be more willing to welcome observers into their classrooms.

Conclusion

Based on my interpretations of the data, I argue that the teachers and evaluators at Desert Middle School have embodied a neoliberal discourse, which has shaped the way in which they regard their teaching practice, quality, and worth. Beyond thinking of themselves in such a way, they also have constructed a dualistic view of what it means to be an acceptable teacher at DMS. Collectively, they have imposed a set of characteristics onto teachers who have either left the school or been terminated. Such teachers have been labeled as being afraid of accountability, confused about the mission, or not really in it for the students. By creating this unacceptable type of teacher, the participants have something against which to compare themselves. In other words, they have used the unacceptable criteria to justify their own subjectivity to the evaluation system. In doing so, they have come to monitor and adjust their conduct, making them responsibilized subjects, or what Brown (2003) called "entrepreneurial actors across all dimensions of their lives," (p. 38).

CHAPTER 6

Conclusions and Implications

In this final chapter, I will present the overall conclusions of the dissertation. To begin, I will provide a brief summary, followed by specific connections and conclusions regarding the three driving research questions of the study. Then I will discuss the challenges I faced and lessons I learned along the way. Finally, I will discuss the implications for policy, practice, and future research.

Summary of the Study

Recent federally funded policy initiatives, such as Race to the Top (RttT) and the Teacher Incentive Fund grants program have created substantial changes in the way teachers are evaluated across the US. For the first time in history, teachers' evaluations are to be based, at least in significant part, on student achievement scores as measured by large-scale standardized assessments. Observation rubrics are also commonly included in evaluations as a means of measuring classroom and professional performance. These evaluation methods have ignited a public debate, garnering the attention of teachers, academics, politicians, think tanks, and the media. Specifically, teachers and education researchers have grown concerned with the ability of the instruments to capture teacher quality reliably, validly, and fairly (Baker et al., 2013; Baker et al., 2010; Berliner, 2013; Hill, Kapitula, & Umland, 2011; Papay, 2010). While there might be less contention regarding observation rubrics, teacher evaluations have been the source of protests (e.g., the Chicago Teacher Strike), lawsuits (Jordan, 2013), and other public debates. Scholars have also written books about the agenda behind such evaluation methods (Amrein-Beardsley, 2014; Berliner & Glass, 2014; Ravitch, 2013). Regardless of the pushback,

nearly all states in the U.S. rely on such methods to evaluate, and make personnel decisions about their teachers (Collins & Amrein-Beardsley, 2014).

Research about the methodological properties of teacher evaluation systems continues to grow; yet missing from the literature has been a policy-as-discourse (Bacchi, 2000) approach that seeks to understand how policies work to constitute both solutions and problems. While traditional policy analyses seek to evaluate either the effectiveness of a policy, the un/intended consequences of the policies, or to investigate the instruments used to carry out such policies, policy-as-discourse analyses flip the focus to look at how the policy works to define the very problem that it attempts to solve. In the case of teacher evaluations, instead of trying to understand how well VAMs capture the construct of teacher quality, the policy-as-discourse analyst might try to understand how the evaluation instruments work to define (or problematize) the construct of teacher quality. That was the driving motivation for this study.

Calling on Foucault's (1984; 1991) governmentality framework (i.e., the strategy of governing the conduct of individuals), I was interested in understanding two distinct, but related, issues: 1) the way in which teacher evaluation policies, practices, and instruments work to problematize teachers and teacher quality; and 2) the way in which teachers at one Arizona middle school have embodied such a discourse to think about and qualify themselves, as well as their teaching practice, quality, and worth. To answer both of these overarching questions, I used a discursive analytical framework and complementary methodological approaches—one that focused on documents, practices, and instruments, and another that focused on teacher interview data.

For the first part of the analysis, I collected official policy documents, political

speeches, press releases, and promotional materials that were relevant and available on the official White House and US Department of Education websites (i.e., whitehouse.gov and ed.gov, respectively). I also collected materials that were directly related to the TAP evaluation system, which is the system in place at Desert Middle School. For this, I collected promotional materials, including speeches, brochures, and other literature as available on their official website (TAPsystem.org). I also attended the 35-hour TAP certification training course where I took field notes and collected the official evaluator handbook and training handbook. First, I analyzed the policy discussions (e.g., speeches, press releases, etc.) to understand how teachers had been positioned as problems within the production function model of schooling. Then I analyzed all of the materials to understand how policies, practices, and instruments were developed and implemented to manage the conduct of teachers in order to minimize their risk to the market-based system.

For the second part of the analysis, I interviewed teachers and their evaluators at one Arizona middle school. The interviews were semi-structured, open-ended interviews. I had a core set of questions that I asked each participant, but I also encouraged the participants to co-construct the interviews with me (Kvale, 1996). This allowed the participants to discuss relevant matters to them that might not have come up otherwise. I interviewed each teacher twice—once in the first half of the school year and once in the second half—and I interviewed the evaluators each once. The participants varied in their subject areas, grade levels, years of experience, backgrounds, and enthusiasm for TAP. When I analyzed the interview data, I was trying to make sense of how the teachers and evaluators have embodied a market-based discourse in terms of how they define

themselves and their conduct, as well as how they qualify their quality and worth based on such measures. In the next sections I will discuss how I brought these two approaches together to answer my overarching research questions.

Teachers as Risky Subjects

In an era of globalization, citizens face an increased pressure to take individual responsibility—to make good, rational choices in preparation for a global competition (Fenwick, 2003). Similarly, traditionally public institutions, like schools, are reconfigured to function as market-based entities. As with any aspect of an economic market, some level of risk is an inherent element. Perhaps the most obvious and concrete example of when education became entrenched in this discourse was in 1983 with the release of *A Nation at Risk*. Public schools were explicitly cast as a threat to the country's economic wellbeing. This had a profound effect on the way in which public schools, and thus administrators, teachers, and students, were positioned in society, which called for new mechanisms of accountability, measurement, and evaluation. Most recently, the focus has narrowed in on teachers.

Teacher evaluation systems that are based on federal initiatives, such as RttT, are consistent with the conceptualization of schools as market-based entities. The findings of this study demonstrate that teachers, specifically, have been positioned as "risky subjects" (Foucault, 1985; McWilliams & Jones, 2005). In each of the discussions about teacher evaluations (e.g., political speeches, official press releases, promotional speeches, etc.) teachers were directly linked with economic values. They were positioned as *the* determining factor of whether or not a student would be economically successful in the future. For example, a report released by the White House, called "Setting the Pace:

Expanding Opportunity for America's Students Under Race to the Top," stated:

> The Race to the Top program has proved that the best and most innovative ideas do not come from Washington. After the program's launch in 2009, dozens of leaders in communities throughout the country answered the call to action and designed new approaches that would better support educators to ensure that students graduate ready for college and careers, enabling students to become productive citizens and out-compete any worker, anywhere in the world, (p. 11).

The excerpt above is a good representation of the way in which the narrative about teachers has (re)constituted educators as the leading source for students' economic trajectory. Not only is the responsibility shifted from "Washington" to teachers, but the students' ability to "become productive citizens and out-compete any worker," can (and should) be "ensure[d]" by the teachers. In this model, the function of the teacher is reduced to the in/ability to prepare students for the market (e.g., add or detract value from the students' future earnings). With this as the established purpose of teachers, then it only makes sense to view teachers as risky, for students have a lot to lose or gain in the process. It also makes sense, then, that techniques be called upon to make sure such risk is minimized. Accordingly, various technologies of governance (Rose, 1999) have been utilized to manage the conduct of teachers. In the following section, I will discuss the specific practices and instruments that are currently being used to carry out such techniques.

Technologies to Manage Teachers' Conduct

> Objects and people are regulated by being represented, described, and formed in a particular conceptual way. Individual subjects are constructed through

'technologies' that make them an object of knowledge. What techniques, Foucault (1977) asks, make an individual 'knowable'? (Fenwick, 2003, p. 340).

Various mechanisms have been put in place at Desert Middle School (and most schools across the country) that attempt to make aspects of teachers and teacher quality into objects of knowledge that can then be acted upon and mangaged. Evaluation practices and instruments, including value-added models (VAMs), observations, rubrics, pre- and post-conferences, self-reflections and assessments, and incentives and punishments, are used to accomplish this task. These mechanisms make possible various technologies of governance that allow teachers to become knowable, and thus measurable, comparable, and evaluate-able.

Numericization of Teachers. The act of attaching standards to teaching only works to constitute teaching as something needing to be regulated (Fenwick, 2003; Nicoll, 1998). Standardizing teaching is one way for teachers and teacher quality to be knowable in terms of numbers, or to be numericized (Rose, 1999). Rose reminds us that "numbers are part of the techniques of objectivity that establish what it is for a decision to be 'disinterested'," (p. 199). As such, decisions about teachers, especially as it relates to qualifying the teacher as good or bad, can be made objectively. However, let us not forget that Rose (1999) also argued that such numbers are subject to the *what* and the *how* something is measured, which has implications for the way in which the object of problematization (i.e., teacher/teaching quality) is constituted by the very instrument meant to measure it (Bacchi, 2000).

Surveillance. Hierarchical surveillance (Foucault, 1977) is another technology that is used to manage the conduct of teachers. Surveillance at Desert Middle School is

done both explicitly and implicitly, both of which require a constant visibility of the teachers. This is accomplished via formal and informal classroom observations—some of which are planned ahead of time and others that are of surprise. This is also accomplished in subtler ways through lesson plan submissions, pre- and post-conferences, and data dashboards (i.e., online portals where administrators can access student test scores and other data). Almost every aspect of teaching, from the external to the internal, is subjected to surveillance, and thus turned into objects of knowledge for the evaluators.

Normalizing Judgments. Standardization, again, constitutes teaching as needing to be monitored. Standardization makes normalizing judgments (Foucault, 1977) possible. Like numericization, normalizing judgment is what allows for objective decisions about teachers and teacher quality to be made. Simultaneously, behaviors that are deemed normal also affect the conduct of teachers in that they adjust their behaviors to comply with the norm. This technique is accomplished in two key ways at DMS. The most prominent way is via observation rubrics that are used for measuring teaching performance during observations. This is also done via self-reflection forms that teachers are required to fill out about themselves and evaluator training methods.

Examination and Audit. By use of the previously discussed technologies, teachers and teacher quality are made objects of knowledge, which makes possible the technologies of examination (Foucault, 1977) and audit (Rose, 1999). Through the collection of artifacts and other pieces of knowledge about the teachers (e.g., VAM scores, rubric scores, etc.), the teacher is turned into a case that can be "described, judged, measured, compared with others, in his very individuality; and it is also the individual who has to be trained or corrected, classified, normalized, excluded, etc.'

(Foucault, 1984, p. 203). At the same time, the teachers internalize these ways of knowing themselves and thus discipline themselves accordingly (Rabinow, 1984).

Teachers' Embodiment of a Market-Based Discourse

> We do not speak a discourse, it speaks us. We are the subjectivities, the voices, the knowledge, the power relations that a discourse constructs and allows. We do not 'know' what we say, we 'are' what we say and do. In these terms we are spoken by policies, we take up the positions constructed for us within policies, (Ball, 1993, p. 14).

In the quotation above, Ball reminds us that discourse is not something that we *do*, but rather something that we *are* and *can be*. In light of the current governing strategy of neoliberalism, teachers are reconfigured as objects of knowledge that can be understood in terms of their market value. Simultaneously, teachers embody this discourse and begin to define and qualify themselves in the same way. This was evident in the interview data of the teacher and evaluator participants at Desert Middle School. The teachers and evaluators demonstrated an embodiment of a neoliberal discourse in two distinct ways (which are also related to the technologies of governance discussed in Chapter 4)—1) the constitution of themselves and their worth in terms of numbers, and 2) their acceptance and need for constant surveillance. These two elements were consistent among all of the participants, though each of their experiences were different in terms of their backgrounds, evaluation scores, and acceptance of TAP.

As for the numericization of themselves, this was coupled with normalizing judgments and an apparent desire to want to fit into the standard norm. This was most evident in the teachers' expressed conflict between doing what was best for their students

versus what was expected of them by the TAP rubric. While the teachers would characterize the rubric in ways such as the "ideal" way of teaching, several of the teachers admitted that the rubric was not best for all of their students, at all times. However, they still qualified teaching in terms of the rubric—as in, they consistently said that the rubric was the best way of teaching, but then they made exceptions for their particular students. This puts the students in a position of deficit, rather than the rubric. Thus begging the question, who or what is the rubric actually good for? In my opinion, the rubric serves a different purpose—one of managing teachers rather than one of helping students learn.

As for the desire for surveillance, this was another common theme across the participants. The teachers discussed two reasons for needing to be observed. First was for improvement—the position that having a mentor teacher, master teacher, or administrator in their rooms telling them what they did well or needed improved was seen as a benefit to professional growth. Another reason teachers and evaluators saw a benefit in having observations was the idea that had there not been any fear of observation, then teachers might not behave in desired, professional ways. In other words, knowing that someone might walk into the teacher's room at any moment kept the teacher in line.

Also, as the participants discussed these two elements in the interviews, the teachers justified their own subjectivities to such practices by comparing themselves against teachers who were not willing and open to such. This most commonly occurred when the participants discussed the teachers who had left DMS since the inception of TAP. These teachers were labeled as being threatened by feedback, scared of being caught, and complacent in their teaching abilities. As the teachers discursively

constructed this idea of the unacceptable teacher, it gave them a binary against which to compare themselves. For example, teachers expressed an acceptance of observation for reasons such as not being afraid of criticism and wanting to improve. The teachers and evaluators shared this common way of looking at observations and evaluations.

Discussion

Ultimately, the findings of this study bring into question the way in which we come to understand complex concepts, such as teacher quality. During the first two decades of the 21st century, we have witnessed a shift in education policy that has thrust teacher quality into the spotlight of the education reform movement. However, the concept of quality is difficult to define and measure because it might mean something different to different stakeholders. For example, some might measure quality in terms of inputs, such as credentials, years, of experience, and the like. On the other hand, some might define quality in terms of outputs, such student scores on standardized achievement tests. In most recent policies (e.g., RttT), outputs have taken precedence in measuring teacher quality via value-added models (, observation rubrics, and other evaluation methods and instruments. Taking a policy-as-discourse approach, I argue that these efforts to measure teacher quality are actually working to define teacher quality in a narrow, market-based way.

Specifically, teacher quality is defined by the way in which is currently being measured by instruments such as VAMs and rubrics. In other words, a teacher is deemed high or low quality as based on that teacher's ability to raise student test scores or to behave in specified ways as per the observation rubric. If the teacher is capable of performing in these ways, then the teacher is labeled as high quality. Other techniques are

also employed in order to encourage the teacher to behave in these desired ways. Teachers and their thinking are under constant surveillance via classroom observations, lesson plan submissions, and conference meetings with evaluators. Teachers are also participants of professional development meetings where they are formally coached on how to behave in accordance with the measurement instruments. Teachers are also encouraged to monitor and adjust their own behavior through self-reflection and self-evaluation via rubrics and self-evaluation forms. All of these techniques work simultaneously to discipline the teachers into behaving in particular ways as deemed by the evaluation system. As such, teacher behavior is being molded to fit the evaluation expectations, rather than the evaluation system working to capture, or measure a natural behavior.

With this in mind, we must question the consequence of defining quality in such a narrow way. Bearing in mind the neoliberal rationality behind contemporary education policies, including the teacher evaluation policies and practices in question, we must consider the guiding principles of market-driven actions. As evident in this study, these might include, but not be limited to: competition, individualism, accountability, standardization, numbers, norms, and market value. Accordingly, what principles are likely to be marginalized so that market-based values can be realized? While the list might be endless, some that be considered are: social justice, equity, compassion, civility, creativity, eco-consciousness, critical thinking, and so on.

The point I want to make here is that, given the current neoliberal governing strategy at play, it seems as though there is a cost to accommodate the demand for everything to be counted, measured, evaluated, and compared. The issue is that only

some things can be counted, and, they can only be counted in certain ways. Thus, if quality must be counted and evaluated, then the way we conceptualize "quality" will always be subject to the way that it is measured. As this study demonstrates, quality has been defined in terms of a teacher's ability to contribute to society in economic terms (either for the country as a whole or for individual students). This was evident in the policy discussions (i.e., teachers as "risky"), as well as the technologies of governance (i.e., numercization, surveillance, normalized judgments, and examination) that have worked together to discipline teachers and minimized their risk to students and society. But, again, is there room for that which cannot be counted and measured in this way?

A Note about Desert Middle School

During this study, I have critiqued the methods by which teachers have been evaluated at Desert Middle School. While I believe this to be of worth to our collective knowledge of teacher evaluation policies, practices, and instruments, I want to also be clear that DMS, specifically, served as a context to understand these matters, and not as the focus of the critique itself. Regardless of level of acceptance for TAP, every participant with whom I spoke shared the same sentiment of DMS that it was a highly supportive and positive place to work. The teachers spoke of the administrators as being attentive, helpful, and overall supportive. The administrators shared an admiration of the teachers as well.

Also, although the evaluation system was held in high regards by most of the participants, when asked what mattered above all else, the principal said the following:

> I'm going to say relationships with kids. Because as long as they have a relationship with kids, of course they have to know their content, but as long as they have a

> relationship with kids, then you can, you can, if they have the kids right here (holds out hand), and the kids know that they like them, you can move forward with that teacher. If there's no relationship, nobody, nothing else matters. They can be brainiac, but if they can't relate to kids then zero is going to happen.

Becky, the principal, was adamant about her position that human judgment mattered at DMS, and I felt the need to include that here.

Implications for Policy and Practice

Teacher evaluation practices that are similar to the ones of this study are currently affecting most teachers in the US (and increasingly other countries as well). While many researchers have focused on concerns related to the methodological issues with VAMs specifically, I have argued that the more commonly recommended practices, such as frequent observations and rubrics (Darling-Hammond, 2013), function in similar ways to VAMs in terms of how they problematize teachers and teacher quality. Taking a policy-as-discourse (Bacchi, 2000) position that policies work to constitute both solutions and problems, I argue that the teacher evaluation policies, practices, and instruments discussed in this study all work similarly to make visible aspects of teaching that are otherwise invisible. In doing so, the way in which teaching quality is defined is a function of the way in which it can be (and has been) measured. Consequently, the policies begin to produce the types of teachers that are measurable by the instruments chosen, thereby eliminating, or at the very least, marginalizing other ways of being a quality teacher.

In other words, the instruments do not solely capture that which already exists; rather, technologies like surveillance and audit work to discipline teachers to behave in the ways designated by the policies/instruments. While it could be argued that the

intention behind the instruments is to produce a particular type of teacher, I want to remind us that the numbers produced by such instruments are always subject to those who choose to measure it and how they choose to measure it (Rose, 1999). Similarly, by producing such types of teachers, at the same time, other ways of being a teacher or other ways of thinking about teacher quality are eliminated. Thus in a mutually constitutive way, quality is reduced to the way in which it can be measured, and at the same time, quality itself actually takes the shape of (or becomes) the expected outcome of the instrument. As an example, VAMs, as of now, can only measure teacher quality in terms of student test scores; thus teacher quality is reduced to student tests, and at the same time, teachers modify their behaviors to fit this expectation. Other scholars have explored this phenomenon from different perspectives and found issues such as narrowed curriculum (Cawelti, 2006; Darling-Hammond, 2007) and teaching to the test (Menken, 2006; Smyth, 2008).

As for what this means for policy and practice is that, while tremendous efforts are focused on trying to find the right tools to capture the construct of teacher quality, we also have a responsibility to realize that the very policies, practices, or instruments that are chosen will simultaneously shape and constitute teacher quality. Put another way, the policies, practices, and instruments meant to measure the construct (of teacher quality), will also work to shape the construct itself. This also eliminates other ways of thinking about, knowing about, or *doing* teacher quality. We must also remember that the way we make sense of and value various social matters is based on a neoliberal discourse that defines everything in terms of market worth. This is a very narrow way of thinking about teacher quality, yet the tools that are currently being used to measure teacher quality are

inline with such a discourse.

Perhaps most importantly, the findings of this study have implications for the way in which we can think about other possibilities of being a quality teacher. With the pinnacle goals of a neoliberal-based teacher quality discourse aiming to increase scores (of both students and teachers), to be more competitive, and to prepare students for a global market, a different set of ultimate goals, such as ones anchored in social justice, equity, and any other possibilities, are not only marginalized, but are likely impossible. Even the title, "Race to the Top," dictates that there will be winners and thus losers. As such, goals of equity and equality are automatically eliminated because there will be losers regardless. With this way of thinking and doing, not only do we eliminate other possibilities of being a quality teacher, but we also eliminate other attainable social goals.

Implications for Policy Research

Through this study, I hope to have aptly demonstrated that ideas, thoughts, and truths get discursively constructed both historically and socially. Teachers at Desert Middle School, who emanated nothing short of love of their school, their students, and their colleagues, have developed a dichotomous view of teachers—ones who are acceptable (e.g., open to audit) and ones who are not (e.g., "scared" of audit). In so doing, they have created a positive culture among themselves, but at what cost? It is not so simple to assume that all teachers who left the school at the beginning of TAP were afraid of being "found out" or did not want to improve. However, the teachers and evaluators have taken up a discourse that limits the possibilities of other types of teachers and their respective qualities. As such, I think this begs the question: what good does a similar approach to policy research have on our collective understanding of policy? In other

words, policy debates, generally speaking, and teacher evaluation policy debates specifically, are often framed in dichotomous ways, which concentrates the blame to a small faction of powerful policymakers and profiteers. So I pose the following questions to the research community:

1. What is the cost of reducing the debate to a dualistic view of good versus bad?
2. How does this framing contribute to a "making up" (Hacking, 1999) of what is good and what is bad?
3. Are other possibilities of analysis and knowledge pushed out in the name of sticking to a dualistic view of policy and knowledge?

To answer these questions, I argue for a more nuanced approach to locating power that avoids a confinement to a definitive group or institution. Related, I argue the same should be applied to thinking about knowledge (Foucault also argued that knowledge and power were inseparable). What I mean is that if we only treat knowledge as a tangible thing that is either good or bad, we might seclude other possibilities for thinking about the object of knowledge. Take teachers for example. The evaluation practices and instruments are intended to turn aspects of teachers into objects of knowledge, which, again, is always subject to the tools available to measure it. As such, I contend that different forms of knowledge should work in complementary, rather than competing ways. On that note, I would like to leave with a comment about Ian Hacking's (1999) quotation on "Making Up People":

> Who we are is not only what we did, do, and will do but also what we might have done and may do. Making up people changes the space of possibilities for personhood. Even the dead are more than their deeds, for we make sense of a

> finished life only within its sphere of former possibilities. But our possibilities, although inexhaustible, are also bounded, (p. 165).

In light of Hacking's words, I hope to have created a small space of opportunity where unknown possibilities about teachers and teacher quality may be imagined, known, and lived.

REFERENCES

Adler, M. (2013). Findings vs. interpretation in "the long-term impacts of teachers" by chetty et al. *Education Policy Analysis Archives, 21*(10), 14.

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

Amrein-Beardsley, A. (2014). *Rethinking value-added models in education: critical perspectives on tests and assessment-based accountability*. New York & London: Routledge.

Amrein-Beardsley, A., & Collins, C. (2012). The SAS Education Value-Added Assessment System (SAS® EVAAS®) in the Houston Independent School District (HISD): Intended and Unintended Consequences. *Education Policy Analysis Archives, 20*(12), 1-36. Retrieved from http://epaa.asu.edu/ojs/article/view/1096

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

Andersen, N. Å. (2003). *Discursive Analytical Strategies: Understanding Foucault, Koselleck, Laclau, Luhmann*. Bristol: Policy Press.

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

Arizona Department of Education (2011). *Arizona framework for measuring educator effectiveness*. Retrieved from: http://www.azed.gov/state-board-education/files/2013/06/arizonaframeworkformeasuringeducatoreffectiveness.pdf

Arizona Department of Education. (2012). *State of Arizona race to the top.* Retrieved from http://www.azed.gov/racetothetop

Arizona Department of Education. U.S. Department of Education, (2012). *State of arizona esea flexibility request*. Retrieved from Ed.gov website: http://www2.ed.gov/policy/eseaflex/approved-requests/az.pdf

Arizona Ready-for-Rigor Project. U.S. Department of Education, Teacher Incentive Fund Application Information. (2010). *Application for new grants under the teacher incentive fund program* (CFDA # 84.385A). Retrieved from http://www2.ed.gov/programs/teacherincentive/apps/a100077.pdf

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

Bacchi, C. (2000). Policy as discourse: What does it mean? Where does it get us? *Discourse: Studies in the Cultural Politics of Education, 21*(1), 45-57.

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

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

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*. Washington, DC: Economic Policy Institute. Retrieved from http://www.epi.org/publications/entry/bp278

Ball, S. J. (1990). *Politics and policy making in education: explorations in policy sociology*. Routledge.

Ball, S. J. (1993). What is policy? Texts, trajectories and toolboxes. *The Australian Journal of Education Studies*, *13*(2), 10-17.

Ball, S. (2003) *Class strategies and the education market: The middle classes and social advantage* (London: Routledge Farmer).

Barnett, J. H., Rinthapol, N., & Hudgens, T. (2014). TAP Research Summary: Examining the Evidence and Impact of TAP: The System for Teacher and Student Advancement. National Institute for Excellence in Teaching.

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

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

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

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

Berliner, D. C., Glass, G. V., (2014). *50 myths & lies that threaten America's public schools: the real crisis in education*. New York: Teachers College Press.

Betebenner, D.W. (2011, April). *Student Growth Percentiles.* National Council on Measurement in Education (NCME) Training Session presented at the Annual Conference of the American Educational Research Association (AERA), New Orleans, LA.

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

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

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

Blunk, M. L. (2007). The QMI: Results from validation and scale-building. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Brandt, C., Mathers, C., Oliva, M., Brown-Sims, M., & Hess, J. (2007). *Examining district guidance to schools on teacher evaluation policies in the midwest region. issues & answers. REL 2007-no. 030.* Regional Educational Laboratory Midwest. 1120 East Diehl Road Suite 200, Naperville, IL 60563.

Brown, W. (2003) Neo-liberalism and the end of liberal democracy, *Theory and Event,* 7(1), 1–43.

Burr, V. (1995) *An Introduction to Social Constructionism* (New York, Routledge).

Capitol Hill Briefing. (2011, September 14). *Getting teacher evaluation right: A challenge for policy makers.* A briefing by E. Haertel, J. Rothstein, A. Amrein-Beardsley, and L. Darling-Hammond. Washington DC: Dirksen Senate Office Building (research in brief). Retrieved from http://www.aera.net/Default.aspx?id=12856

Castellano, K.E. & Ho, A.D. (2013). *A Practitioner's Guide to Growth Models.* Council of Chief State School Officers.

Cawelti, G. (2006). The side effects of NCLB. *Educational Leadership, 64*(3), 64-68.

The Center for High Impact Philanthropy. (2010). High impact philanthropy to improve teaching quality in the u.s. *Blueprint*, Retrieved from http://www.impact.upenn.edu/images/uploads/UPenn_CHIP_TQProjectBlueprint_Mar10(1).pdf

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. *(NBER working paper no. 17699*.) Cambridge, MA: National Bureau of Economic Research. Retrieved from http://obs.rc.fas.harvard.edu/chetty/w19423.pdf

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

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

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

Consolidated Appropriations Act of 2010 (2009). Public Law 111-117. Retrieved from: http://www.gpo.gov/fdsys/pkg/PLAW-111publ117/pdf/PLAW-111publ117.pdf

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

Corbin, J., & Strauss, A. (Eds.). (1997). *Grounded theory in practice*. Sage.

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

Daley, G., & Kim, L. (2012). Retrieved from http://www.tapsystem.org/publications/tap_research_summary_0210.pdf

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

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, Va: Association for Supervision and Curriculum Development.

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

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

Darling-Hammond, D. (2013). *Getting teacher evaluation right: What really matters for effectiveness and improvement*. Teachers College Press, New York, NY.

Davies, B., & Bansel, P. (2010). Governmentality and academic work: shaping the hearts and minds of academic workers. *Journal of curriculum theorizing*, *26*(3).

Dean, M. (1994) *Critical and Effective Histories: Foucault's Methods and Historical Sociology* (London: Routledge).

Dean, M. (1999). *Governmentalty: Power and rule in modern society*. London: SAGE.

Dey, I. (1993). What is qualitative analysis. *Qualitative data analysis*, 31-54.

Duncan, A. (2010). Race to the top – integrity and transparency drive the process [Web log message]. Retrieved from http://www.ed.gov/blog/2010/01/race-to-the-top-–-integrity-and-transparency-drive-the-process/

Duncan, A. U.S. Department of Education, (2010). *Reform, accountability, and leading from the local level: Secretary arne duncan's remarks to the national league of cities' congressional city conference*. Retrieved from website: http://www.ed.gov/news/speeches/reform-accountability-and-leading-local-level-secretary-arne-duncans-remarks-national-

Eisenhower, D. (1958). *State of the union address*. Retrieved from website: http://www.eisenhower.archives.gov/all_about_ike/speeches.html

Ellett, C. D., & Garland, J. (1987). Teacher evaluation practices in our 100 largest school districts: Are they living up to "state-of-the-art" systems? *Journal of Personnel Evaluation in Education*, 1, 69–92.

Fairclough, N. (1989). *Language and power*. London: Longman.

Fairclough, N., & Wodak, R. (1997). Critical discourse analysis. In T. van Dijk (Ed.), *Discourse as social interaction* (pp. 258-284). London: Sage.

Fenwick, T. (2003) The 'good' teacher in a neo-liberal risk society: a Foucaultian analysis of professional growth plans. *Journal of Curriculum Studies*, 35:3, 335-354. DOI:

10.1080/00220270210151089

Foucault, M. (1977). *Discipline and punish: The birth of the prison*, London: Allen Lane.

Foucault, M. (1980). *Power/knowledge: Selected interviews and other writings*, 1972-1977, ed. and trans. C. Gordon (New York: Pantheon Press).

Foucault, M. (1982). Is it really important to think? *Philosophy and Social Criticism*, *9*(1), 29–40. http://dx.doi.org/10.1177/019145378200900102

Foucault, M. (1984). The means of correct training. In P. Rabinow (ed.), *The Foucault Reader* (New York: Pantheon), 188–205.

Foucault, M. (1985). *The use of pleasure: The history of sexuality (vol. 2)*. R. Hurley, Trans. London, Penguin.

Foucault, M. (1991) Governmentality. In G. Burchell, C. Gordon and P. Miller (eds), *The Foucault Effect: Studies in Governmentality* (Chicago: University of Chicago Press), 87–104.

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

Glazerman, S., & Seifullah, A. (2012). *An evaluation of the Chicago teacher advancement program (Chicago TAP) after four years. final report* Mathematica Policy Research, P.O. Box 2393, Princeton, NJ.

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

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

Goldstein, J., & Noguera, P. A. (2006). A thoughtful approach to teacher evaluation. *Educational Leadership, 63*(6), 31-37.

Graham, C., & Neu, D. (2004). Standardized testing and the construction of governable persons. *Journal of Curriculum Studies*, 36(3), 295-319.

Gramsci, A. (1971). *Selections from the Prison Notebooks of Antonio Gramsci: Ed. and Transl. by Quintin Hoare and Geoffrey Nowell Smith*. G. Nowell-Smith, & Q. Hoare (Eds.). International Publishers.

Guba, E. G. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *ECTJ*, *29*(2), 75-91.

Guskey, T. R. (2002). Does it make a difference?: Evaluating professional development. *Educational Leadership,* 59(6), 46-51.

Hacking, I. (1999). Making up people. *The science studies reader*, *18*, 590.

Hacking, I. (2004). Between Michel Foucault and Erving Goffman: between discourse in the abstract and face-to-face interaction. *Economy and Society*, *33*(3), 277-302.

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

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

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

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

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

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

Hill, H. C., Charalambous, C. Y., Blazar, D., McGinn, D., Kraft, M. A., Beisiegel, M., . . . Lynch, K. (2012). Validating arguments for observational instruments: Attending to multiple sources of variation. *Educational Assessment, 17*(2), 19.

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

Hodkinson, P. (2008). Scientific research, educational policy, and educational practice in the United Kingdom: The impact of the audit culture on further education. *Cultural Studies↔ Critical Methodologies*, *8*(3), 302-324.

Holloway-Libell, J., & Collins, C. (2014). VAM-Based teacher evaluation policies: Ideological foundations, policy mechanisms, and implications. *InterActions: UCLA Journal of Education and Information Studies*, *10*(1).

Hudson, S. (2010). The effects of performance-based teacher pay on student achievement: Discussion paper. *Stanford Institute For Economic Policy Research*, Retrieved from http://www.stanford.edu/group/siepr/cgi-bin/siepr/?q=system/files/shared/pubs/papers/09-023_Paper_Hudson.pdf

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

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

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

Jordan, G. (2013, April 16). Teachers union files federal lawsuit challenging Florida teacher evaluations. *StateImpact*. Retrieved from http://stateimpact.npr.org/florida/2013/04/16/teachers-union-files-federal-lawsuit-challenging-florida-teacher-evaluations/

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

Kennedy, M. M. (2010). Attribution error and the quest for teacher quality. *Educational Researcher, 39*(8), 591-598. doi:10.3102/0013189X10390804

Kersting, N. B., Chen, M., & Stigler, J. W. (2013). Value-added added teacher estimates as part of teacher evaluations: Exploring the effects of data and model specifications on the stability of teacher value-added scores. *Education Policy Analysis Archives, 21*(7), 1-39. Retrieved from http://epaa.asu.edu/ojs/article/view/1167

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). What do test scores in Texas tell us? *Education Policy Analysis Archives, 8*(49), 1-22. Retrieved from http://epaa.asu.edu/epaa/v8n49

Koretz, D. (1996). Using student assessments for educational accountability. In E. A. Hanushek & D. W. Jorgenson (Eds.), *Improving America's schools: The role of incentives.* Washington, DC: National Academy Press.

Kvale, S. (1996). The interview situation. *Interviews. An Introduction to Qualitative Research Interviewing*, 124-143.

Lemke, T. (2002). Foucault, governmentality, and critique. *Rethinking marxism*, *14*(3), 49-64.

Linn, R L., & Haug, C. (2002). Stability of school-building accountability scores and gains. *Educational Evaluation and Policy Analysis, 24*, 29-36. doi:10.3102/01623737024001029

Lockwood, J. R. & McCaffrey, D. F. (2009). Exploring student-teacher interactions in longitudinal achievement data. *Education Finance and Policy, 4*(4), p. 439-467. doi:10.1162/edfp.2009.4.4.439

Loup, K. S., Garland, J. S., Ellett, C. D., & Rugutt, J. K. (1996). Ten years later: Findings from a replication of a study of teacher evaluation practices in our 100 largest school districts. *Journal of Personnel Evaluation in Education*, 10(3), 203–226.

Lyotard, J. F. (1999). The postmodern condition. *Modernity: Critical Concepts*, *4*, 161-177.

Marshall, J. D. (Ed.). (2004). *Poststructuralism, philosophy, pedagogy*. Kluwer Academic Publishers.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A. & Hamilton, L. (2004). Let's see more empirical studies on value-added modeling of teacher effects: A reply to Raudenbush, Rubin, Stuart and Zanutto, and Reckase. *Journal of Educational and Behavioral Statistics*, *29*(1), 139-143. doi:10.3102/10769986029001139

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

McWilliam, E. (2002). Against Professional Development. *Educational Philosophy & Theory*,*34*(3), 289-299. doi:10.1080/00131850220150246

McWilliam, E., & Jones, A. (2005). An unprotected species? On teachers as risky subjects. *British Educational Research Journal*, 31(1), 109-120.

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

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

Nicoll, K. (1998) 'Fixing' the 'facts': flexible learning as policy invention. *Higher Education Research And Development,* 17 (3), 291-304.

NIET. (2006). *Tap system leadership handbook*. Retrieved from http://www.tapsystemtraining.org/Portals/0/TAPHandbook.pdf

NIET. (2013). *Tap system training portal*. Retrieved from http://tapsystemtraining.org/Default.aspx?alias=www.tapsystemtraining.org

Obama, B. U.S. Department of Education, Office of the Press Secretary. (2009). *Remarks by the president to the Hispanic chamber of commerce on a complete and competitive American education*. Retrieved from website: http://www.whitehouse.gov/the_press_office/Remarks-of-the-President-to-the-United-States-Hispanic-Chamber-of-Commerce/

Oliva, M., Mathers, C., & Laine, S. (2009). *Effective evaluation* National Association of Secondary School Principals. 1904 Association Drive, Reston, VA 20191-1537.

O'Malley, K. J., Moran, B. J., Haidet, P., Seidel, C. L., Schneider, V., Morgan, R. O., Kelly, P. A., Richards B. (2003). Validation of an observation instrument for measuring student engagement in health professions settings. *Evaluation and the Health Profression*, 26(1), 86-103.

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

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

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

Peters, M. (1996). *Poststructuralism, politics, and education*. Westport, Conn.: Bergin & Garvey.

Peters, M. (2009). Education, enterprise culture and the entrepreneurial self: A Foucauldian perspective. *The Journal of Educational Enquiry*, *2*(2).

Power, M. (1997). *The audit society: Rituals of verification.* Oxford, UK: Oxford University Press.

Prewitt, K. (1987). 'Public statistics and democratic politics' in Alonso and Starr 1987, pp. 261-74.

Rabinow, P. (1984) Introduction. In P. Rabinow (ed.), *The Foucault Reader* (New York: Pantheon Books), 3–29.

Rabinow, P. (Ed.). (1991). *The foucault reader* (p. 244). London: Penguin.

Rabinow, P., & Rose, N. (2003). Foucault today. *P. Rabinow & N. Rose (Eds.(2003) The essential Foucault (1954-1984)*.

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

Ransom, J. (1997). *Foucault's discipline: The politics of subjectivity*. Duke University Press.

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

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

Ravitch, D. (2013). *Reign of error: the hoax of the privatization movement and the danger to America's public schools*. Knopf.

Recovery and Reinvestment Act of 2009 (2009). Division A, Title VIII, Public Law No. 111-5. Retrieved from: http://www.gpo.gov/fdsys/pkg/PLAW-111publ5/pdf/PLAW-111publ5.pdf

Rogers, R., Malancharuvil-Berkes, E., Mosley, M., Hui, D., & O'Garro, G. J. (2005). Critical discourse analysis in education: A Review of the literature. *Review of Educational Research*, 75(3), 365-416.

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

Rose N (1996) Governing 'Advanced' Liberal Democracies. In A Barry, T Osborne, & Rose (eds) *Foucault and Political Reason: Liberalism, Neo-liberalism and Rationalities of Government.* London: UCL Press.

Rose, N. (1999). *Powers of freedom: Reframing political thought*. Cambridge university press.

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

Robertson, R. (1990). Mapping the global condition: Globalization as the central concept. *Theory, Culture and Society*, *7*(2-3), 15-30.

Rose, N. (1999). *Powers of freedom reframing political thought*. Cambridge, United

Kingdom: Cambridge University Press.

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

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

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

Ruby, J. (1980). Exposing yourself: reflexivity, anthropology, and film.*Semiotica*, *30*(1-2), 153-180.

Saldaña, J. (2013). *The coding manual for qualitative researchers*. London: Sage.

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

Saul, J. R. (2005) *The collapse of globalism and the reinvention of the world* (Camberwell, Viking).

Schacter, J., & Thum, Y. M. (2005). TAPping into high quality teachers: Preliminary results from the teacher advancement program comprehensive school reform. *School Effectiveness and School Improvement, 16*(3), 327-353. doi:http://dx.doi.org/10.1080/13803610500146202

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

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

Simon, A., & Boyer, E. G. (1969). Mirrors for behavior, *An anthology of classroom observation instruments.* ERIC document Reproduction No. 031613.

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

Seidman, I. (2013). *Interviewing as qualitative research: A guide for researchers in education and the social sciences*. Thousand Oaks, CA: Sage.

Spradley, J. P. The ethnographic interview, 1979. *New York: Holt, Rinehart and Winston*.

Starr, P. (1987). 'The sociology of official statistics' in Alonso and Starr 1987, pp. 7-58.

Stake, R. E., & Trumbull, D. J. (1982). 1 Naturalistic Generalizations.

State of Arizona Senate (2010). Senate Bill 1040 (A.R.S. §15-203(A)(38)). Retrieved from: http://www.azleg.gov/legtext/49leg/2r/bills/sb1040h.pdf

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

Stiggans, R. J., & Duke, D. L. (1988). *The case for commitment to teacher growth: Research on teacher evaluation*. Albany, NY: State University of New York Press.

Sweeney, J., & Manatt, D. (1984). *A team approach to supervising the marginal teacher.*

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

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

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

U.S. Department of Education. (2009). *Race to the top program executive summary*. Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

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

Van Tassel-Baska, J., Quek, C., & Feng, A. (2007). The Development and Use of a Structured Teacher Observation Scale to Assess Differentiated Best Practice. *Roeper Review, 29(2), 84-92*

van Dijk, T. (1993). *Principles of critical discourse analysis*. Discourse & Society, 4(2), 249-283.

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

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

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

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

APPENDIX A

DATA INCLUDED IN PART 1 ANALYSIS:

POLICIES, PRACTICES, AND INSTRUMENTS

Data	Date	Type
Race to the Top (RttT)	2009	Official Policy Document
Elementary and Secondary Education Act (ESEA) Flexibility (i.e., No Child Left Behind Waiver)	September, 2011	Official Policy Document
Teacher Incentive Fund (TIF) grants program	November, 2010	Official Policy Document
Arizona's RttT Application	December 13, 2011	Official Policy Document
Arizona's ESEA Waiver application	February 28, 2012	Official Policy Document
Arizona Ready-for-Rigor Application (TIF grant application)	November 30, 2010	Official Policy Document
Arizona's Framework for Measuring Educator Effectiveness	April 25, 2011	Official Policy Document
Letter to Chief State School Officers regarding ESEA waiver extension	November 14, 2013	Official policy letter
Letter to Chief State School Officers regarding ESEA waivers	June 18, 2013	Official policy letter
RttT Phase 3 Guidance and Frequently Asked Questions	October 27, 2011	Policy support materials
RttT Program Guidance and Frequently Asked Questions	May 27, 2010	Policy support materials
Summary of Considerations to Strengthen State Requests for ESEA Flexibility	Not provided	Policy support materials
Building or Buying Assessments to Measure Student Growth (webinar)	May 2013	Policy support materials
Use of School-Wide Growth	April 26, 2013	Policy support materials
Building Evaluation Systems that Support Educators of Students with Disabilities (webinar)	June 12, 2013	Policy support materials

Forum on ESEA Flexibility	September 29-30, 2011	Policy support materials (transcript)
RttT Expansion	January 19, 2010	Press Release
18 States and D.C. Named as Finalists for RttT	July 27, 2010	Press release
Education Department Announces 16 Winner of RttT-District Competition	December 11, 2012	Press release
President Obama, U.S. Secretary of Education Duncan Announce National Competition to Advance School Reform Obama Administration Starts $4.35 Billion "Race to the Top" Competition, Pledges a Total of $10 Billion for Reforms	July 24, 2009	Press release
States Continue Progress During Second Year of RttT	February 1, 2013	Press release
16 Finalists Announced in Phase 1 of RttT Competition Finalists to Present in Mid-March; Winners Announced in Early April	March 4, 2010	Press release
U.S. Department of Education Opens RttT Competition	November 12, 2009	Press Release
President Obama, U.S. Secretary of Education Duncan Announce National Competition to Advance School Reform	July 24, 2009	Press Release
Secretary Duncan Challenges National Education Association to Accelerate School Reforms	July 2, 2009	Press release
Duncan Offers Stimulus Funds for States to Develop Rigorous Assessments Linked to Common Standards	June 15, 2009	Press release
U.S. Secretary of Education Calls on State Officials and Researchers to Deliver Honest Answers about Reforms	June 8, 2009	Press release
States Open to Charters Start Fast in ‘RttT’	June 8, 2009	Press release
Secretary Duncan Asks: Will California Lead or Retreat in Public K-12 Education’s RttT?	May 22, 2009	Press release

Secretary Duncan Sets Tone for 'RttT' by Naming Innovative New Leader	May 19, 2009	Press release
Obama Administration Approves NCLB Flexibility Requests for Delaware, Georgia, Minnesota, New York and South Carolina	July 31, 2014	Press release
U.S. Department of Education Approves Extensions for States Granted Flexibility from NCLB	July 3, 2014	Press release
States Granted Waivers from NCLB Allowed to Reapply for Renewal for 2014 and 2015 School Years	August 29, 2013	Press release
NCLB: Early Lessons from State Flexibility Waivers	February 7, 2013	Press release
Graduation Rates and ESEA Flexibility	Not provided	Press release
U.S. Department of Education Boosts District-Led Efforts to Recognize and Reward Great Teachers and Principals Through the 2012 Teacher Incentive Fund	September 27, 2012	Press release
2012 Teacher Incentive Fund Invites Districts to Pursue a New Vision for Human Capital Through Stronger Evaluations and Greater Professional Opportunities	June 8, 2012	Press release
U.S. Department of Education Announces $442 Million in Teacher Incentive Fund Grants; 62 Winners from 27 States	September 23, 2010	Press release
Department Begins Competition for $437 Million in Teacher Incentive Fund Grants	May 20, 2010	Press release
Remarks by the President on Race to the Top at Graham Road Elementary School	January 19, 2010	Speech
Address by the Secretary of Education To the National Education Association	July 2, 2009	Speech
The Obama Record in Education—Secretary Duncan's Remarks to the Mom Congress	April 30, 2012	Speech

Moving Forward, Staying Focused—Remarks of Arne Duncan, National Press Club	October 2, 2012	Speech
Change is Hard—Remarks of U.S. Secretary of Education Arne Duncan at Baltimore County Teachers Convening	August 22, 2012	Speech
The Quiet Revolution: Secretary Arne Duncan's Remarks at the National Press Club	July 27, 2010	Speech
"A Message from the Founder" (Lowell Milken)	Not provided	TAP materials
14th National TAP Conference Develops Teacher Leaders		TAP materials
Introductory Remarks from the National Governors Association, 97th Annual Meeting, July 17, 2005, Lowell Milken	July 17, 2005	TAP materials
Video on Teacher Quality Crisis (TAP)	Not provided	TAP materials
TAP response to the Teacher Quality Crisis	Not provided	TAP materials
TAP System website		TAP materials
TAP Evaluator Certification Course--24 hours (field notes)	June 3-6, 2013	TAP materials
Education is the Answer (Lowell)	May 2, 1996	TAP materials
TAP Evaluator Handbook	June 2013	TAP materials
TAP Evaluator Training Manual	June 2013	TAP materials

APPENDIX B

ANALYTIC MEMOS

The following memos were collected between the dates of October 2013 and August 2014. These were my ongoing notes, thoughts, and questions as I collected, analyzed, and made sense of the data.

An evaluation process chooses what is being evaluated. But the "what" is always limited by the "how" and the "how" shapes the behavior of the evaluatees

At first I saw contradictions as sources of interest in that I didn't know why or how one person could hold such conflicting thoughts. But then I started to realize that, since participants' responses did not represent objective reality, then those "contradictions" weren't necessarily confusing or conflicting per se. Their responses were situational and contextual. As such, the goal became to better understand the conditions that must be present in order for these various responses to exist simultaneously. This also went for contradictions between participants—each participant's "reality" is fair and real to that participant. The goal is not to figure out who was closer to some real "truth."

—"To enquire into this transformation of difficulties into problems which demand solutions is not to arbitrate between existing responses, but to 'free up' possibilities. The act of thinking is an act of modal transformation from the constative to the subjunctive, from the necessary to the contingent," (Rainbow & Rose, 2003, p. 13).

Not framing this as an idealogical argument (not employing a critical framework)

How do TAP-based policies and systems encourage (right word?) teachers and administrators to govern themselves?

What knowledge is available to the teachers/administrators that might influence the way they view quality teaching and relate themselves against that goal?

a. What has been constituted as knowledge? What sources have contributed to that knowledge?

"I think it just gives us a tool to kind of (*) discuss those things, um, but also just have very observable actions associated with it, so not give better feedback, but I want to see KIDS giving feedback to each other. I want to see kids using that feedback to adjust that learning. And so it's really targeted, um, in terms of what we're looking for and how we can move teachers."

- This statement is really interesting to me. Kind of a chicken or egg dilemma. She says that the rubric targets what she "want[s] to see." But doesn't she "want to see" it because that's what the rubric asks for?? Did she really want that before the school adopted the rubric? What "knowledge" is she pulling from that legitimizes this practice??

The consciously competent teacher: "Yeah, and, I think, and like I think one of the things that we learned with TAP, which was interesting, was the idea of like consciously competent, so someone who knows what they're doing and they're conscious of the decisions they're making. And I feel like our skilled teachers are at that level, so they can justify, you know what, yeah, I should have differentiated, but I didn't and here is the exact reasons why."

"my consciously competent teachers are telling the kids too. It's very clear in the lesson.

They'll say, we're going to work in these groups today because I want you to do blah blah blah, so the kids are aware of it, the observer is aware of it, and it's clear that the teacher is, so. I think (**) there's a difference, and so, the evidence in the classroom really shows it, and then in the lesson plan too, you'll see what they're taking note of versus what they're recording just for the sake of recording and what they're recording because they want you to know the decisions they've made and why they've made them."

- Making their thought process explicit and available to observer/evaluator (is this surveillance? Examination?).

It's like this ongoing dance to find this middle ground so that as evaluators and evaluatees get to a place where they are sustainable and agreeable.
People prefer the unannounced. –why?
Further research: what are their repertoires of knowledge? Where do these come from, and how are they discursively constructed over time?
The teachers are asked to self-reflect…but they see themselves more wholly than an observer would ever do. The teacher thinks about her/his behaviors of that lesson, that day, that week, etc., while the observer can only think about the one lesson on the one day.
There seems to be a disconnect between the student growth and evaluation. "Evaluation" seems to trigger "observation evaluation" rather than the SGP component. The SGP component seems to be an afterthought or a "it is what it is"
There's an ever-ending quest for rigor—wasting time trying to figure out how something fits the rubric.
This study is less about evaluating the way in which TAP works in a district and more about how we can use theory to think about how individuals within an organization consume (and produce) knowledge
Leaving is regarded as a sign of can't make it, or doesn't want to be held accountable.
"This is a judgment on an individual lesson" —like the fact that growth scores are a judgment of an individual test
—Effective teacher? Or effective at teaching the lessons constructed/encouraged by TAP?

Successful vs effective
Understanding this system requires thinking about knowledge and knowledge production. A teacher evaluation system is built on particular assumptions about knowledge and knowledge production. How do we "know" that a teacher is good? Where do the ideas about how we "know" this come from? These assumptions are predicated on ways of thinking and knowing (i.e., discourses).
"My belief is that if those frameworks are made visible, possibilities may open up to rethink the conceptions of teaching and teacher education onto different paths," (Friedrich, 2014).
Need to analyze "accountability" in terms of risk. How does accountability help to manage risk? How is this discursively constructed?
Trigger moments (aha moments)
6. When teachers were happy with the system, but said that it wasn't good for all students. The teachers saw themselves in relation to the system, not necessarily in

relation to their students.

7. Many ways to be a teacher (different frameworks)
8. A need to be watched/observed (I won't care unless someone is in my room). This doesn't follow other lines of reasoning (e.g., I'm a career teacher; I love teaching, etc.).
9. Given my theoretical framework, these inconsistencies did not represent lies, confusion, or conflicting cognitions, rather they called for a more nuanced analysis of the conditions that potentially made these inconsistencies possible. It was at that moment of realization that my dissertation took a new turn and my interview data became but one of several elements of analysis, for I realized that in order to understand how these teachers see themselves within the system, I must attempt to understand how the system simultaneously works to "make up" effective teachers.

All of these "things" don't neatly trace back to one entity (e.g., right-wing conservatives, Breaking the mold of the powerful (e.g., manipulative politicians) and victims (e.g., teachers)

"utilized a range of techniques that would enable the state to divest itself of many of its obligations, devolving those to quasi-autonomous entities that would be governed at a distance by means of budgets, audits, standards, benchmarks, and other technologies that were both autonomizing and responsibilizing," (Rose et al., 2006, p. 91).

—Rubrics, conferences, observations, self-reflections, etc.

Teachers problematized

Tools/instruments (why quantifying instruments "make sense")

Concepts:

- Risk
- Examination
- Surveillance

This is a critique of [VAMs], but instead of within a typical policy framework (i.e., does this policy work?), I argue that VAMs discursively construct a certain way of being a teacher, and that teachers begin to think about themselves in those terms and behave accordingly.

Since the students' "achievement" can only be measured on math/reading, the students are pulled from electives to have extra math/reading time.

Time is a function of the system—time that can be spent increasing scores is important

The following quote demonstrates how this particular teacher disapproves of testing not because art is immeasurable, but because teacher observation of quality art is not good enough. She considers art and the other electives "filler[s]," reasoning that personal teacher observation is not "concrete data" and that "just seeing" does not really say anything about how well the students have done.

> "So we choose to give them this test to then, well like this year would be just a year to see if we wanted to do it, like just to test it out, test out the test. And um, I think almost unanimously, we all were like, this is ridiculous because the kids are tested

> SO much, and now we're going to pull them to be tested for our areas too? And then like half of the test still said, well it's just observation based on your own personal observation, so then it's still not like concrete data necessarily. Like did they improve or did they not improve, or portfolio-based, like just seeing. And because of the fact that we don't have them for a full year, we only have them for a trimester, they rotate, so it's like, AND they mix the seventh and eighth graders. So then it's like you can't even do a seventh grade curriculum and then an eighth grade curriculum, it's just so (*) once again, the fact that they have the arts and the electives here is awesome. How it's done is still, it's just like you're a filler kind of thing."

The analysis of the teacher transcripts will be done by case – keep them in context
Maybe evaluation can't really exist? Evaluation assumes that it is possible to capture something that exists in reality and is untouched by the evaluative tool. However, the "thing" in question will always be dictated by the tool designed to measure it.
Can quality naturally exist, or is it always produced? Is there a way of knowing?

APPENDIX C

DESCRIPTIVE CODES FROM FIRST ROUND OF CODING

Accountability
Adjusted Instruction
Administrators
AIMS
Background
Cluster
Coaching
Collaboration
Conferences
Culture
Evaluation Process
Evaluator-Teacher Relationship
Feedback
Future Plans
Good Teaching
Leadership Team
Master Teacher
Mentor Teacher
Money
New Teachers
NWEA
Observation Performance
Observations
Osborn
Planning
Refinement/Reinforcement
Resistance
Rubric-based Evaluation
Scores
SGP-based Evaluation
SGP-SKR Relationship
Standards
Students
TAP as a System
TAP as development
TAP Rubric
Teacher as Evaluation Number
TFA vs Traditional
Transition into TAP
Unannounced
Years teaching