THE RESPONSE OF CENTRAL OFFICE LEADERS TO THE TEACHER EFFECTIVENESS MODEL

A Dissertation in Education Leadership

by

Andrew E. McCrea

Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

December 2015
The dissertation of Andrew E. McCrea was reviewed and approved* by the following:

Nona A. Prestine
Professor of Education (Educational Leadership)
Dissertation Advisor
Chair of Committee

Edward J. Fuller
Associate Professor of Education (Educational Leadership)

Ronald Musoleno
Associate Professor of Education Policy Studies

Edgar Yoder
Professor of Agriculture and Extension Education

Kia Schafft
Associate Professor of Education (Education Leadership) and Rural Sociology
Graduate Program Chair Education Leadership

*Signatures are on file in the Graduate School.
ABSTRACT

The current study extends the research on teacher evaluation models beyond psychometric reliability and validity by examining the role central office staff have in implementing reform. Central office staff are often the first educators exposed to new reform-related information and play an active role in implementation by interpreting what they learn and shaping it to fit their local context. The literature base suggests that background knowledge and social interaction are two factors that influence how interpretations are made. These interpretations shape reform at the local level, and thus, affect whether new practices are aligned only with surface-level behaviors of change or deeper shifts in values, beliefs, and goals. This study considered how background knowledge and social interaction, as well as other factors like the perceived value or worth of reform, competing issues in a district, and personal beliefs of central office staff influenced the central office staff response to new reform oriented information from Pennsylvania’s Teacher Effectiveness Model (TEM). Results found that factors related to personal beliefs, the perception of value or benefit a reform is perceived to have in a district, and the social network of central office staff interacted to shape the sense that was ultimately made. Findings also indicated a tendency for districts to leverage the TEM as a way of focusing on pre-existing local initiatives. Personal beliefs and perception of value or benefit are factors that have received limited attention in the sensemaking literature, as has the complex interaction between these and other factors. Current findings suggested additional attention should be given to these areas.
# Table of Contents

List of Tables.............................................................................................................vii
List of Figures.............................................................................................................viii
Acknowledgements.................................................................................................ix

Chapter 1. INTRODUCTION.......................................................................................1
   Background of the Problem.........................................................................................1
   Pennsylvania’s Teacher Effectiveness Model............................................................2
   Reliability and validity of educator evaluation models..........................................3
   Role of the central office in reform implementation............................................4
   Theoretical Frame.....................................................................................................6
   Purpose of the Study & Research Question............................................................7
   Research Design.......................................................................................................8

Chapter 2. LITERATURE REVIEW..........................................................................10
   Teacher Evaluation Background............................................................................10
   Background to Pennsylvania’s Teacher Effectiveness Model.................................12
   Components of Pennsylvania’s Educator Effectiveness Model..............................16
      Observation...........................................................................................................18
      Student Outcome Data........................................................................................20
         Pennsylvania’s School Performance Profile....................................................22
         Pennsylvania’s Value Added Model.................................................................23
         Pennsylvania’s Student Learning Objectives..................................................25
      PA Model Summary............................................................................................27
   Reliability and Validity of Model Components.......................................................28
      Teacher Observation.............................................................................................30
      School Metrics......................................................................................................32
      Value Added Modeling........................................................................................36
         Statistical Assumptions.....................................................................................39
         Concerns.............................................................................................................40
      Student Learning Objective Research...............................................................42
      Consequence Validity...........................................................................................43
      Reliability and Validity Research Summary.......................................................44
   Role of Central Office Staff in Reform Implementation...........................................45
      First or Second Order Change..........................................................................47
         Capacity............................................................................................................48
         Resistance.........................................................................................................48
      Role of Central Office Summary.......................................................................50
   Sensemaking as a Theoretical Frame....................................................................51
      Background Knowledge.......................................................................................54
      Social Processes..................................................................................................56
      Other..................................................................................................................57
   Literature Review Summary...................................................................................60
Chapter 3. METHODOLOGY

- Qualitative Rationale
- Meaning and Understanding
- Researcher as Primary Instrument
- Inductive Process
- Thick and Rich Description

Case Study Rationale
Site and Sample Selection
Research Strategies
Data Collection & Analysis
Reliability and Validity
Limitations and the Role of the Researcher

Chapter 4. RESULTS

District A
- Background
- Perceptions of Value or Worth
- Competing Factors
- Past Experiences
- Personal Beliefs
- Social Interactions
- Interaction of Factors in District A

District B
- Background
- Perceptions of Value or Worth
- Competing Factors
- Past Experiences
- Personal Beliefs
- Social Interactions
- Interaction of Factors in District B

District C
- Background
- Personal Beliefs
- Perceptions of Value or Worth
- Competing Factors
- Social Interactions
- Interaction of Factors in District C

District D
- Background
- Past Experiences & Personal Beliefs
- Perceptions of Value or Worth
- Competing Factors
- Social Interactions
- Interaction of Factors in District D
List of Tables

Table 1: District Demographics.................................................................140
List of Figures

Figure 1: Framework of Factors that Lead to Sensemaking

.................................................. 158
Acknowledgements

I would like to thank the members of my committee for their guidance during this process. I would specifically like to acknowledge my chair, Dr. Nona A. Prestine. Her knowledge, experience, patience, and mentorship were invaluable; and have left a lasting impression.
Chapter 1: Introduction

Education policy is marked by a focus on evaluating the local education context in order to ensure accountability for student learning. Research on the accountability models that stem from such policy have focused mainly on technical properties related to reliability and validity, implicitly suggesting that systems will work when they are built to consistently and accurately measure the construct of interest (such as teacher effectiveness). Limited study has considered the extent to which variables beyond that of technical adequacy might contribute to predicting the likelihood of successful accountability-based model implementation. Given the predominate role of central office staff in the policy implementation process, investigating how these leaders affect the success of accountability models is justified. The current study contributes to this investigation by investigating how central office staff made sense of one particular model that resulted from accountability policy.

Background of the Problem

In an effort to increase student academic outcomes, education policy has shifted toward increased amounts of accountability over recent decades (McDermott, 2007). Policy mandates have manifested in school based report cards and other reward and sanction based systems rooted in student performance results (Hanushek & Raymond, 2005; Kane & Staiger, 2002; Ladd, 2001). Such systems have a history at the state level dating back to the latter parts of the twentieth century; but were federalized in 2002 through the No Child Left Behind Act [NCLB] (2002) and the concept of Adequate Yearly Progress (AYP). Adequate Yearly Progress was a national attempt to ensure that all students were achieving at proficient levels. Unfortunately, and similar to earlier state attempts (McDermott, 2007), the mandates of NCLB (2002) including
AYP were limited in regard to reaching stated goals (Koretz, 2009; Jacob & Levitt, 2003; Johnson, 2011; Neal & Schanzenback, 2010).

While federal policy came up short in achieving its goals, legislators and government actors remained steadfast in their desired outcomes. As a result, in 2012, changes to NCLB (2002) were offered to states through a waiver process. Federal waiver principals focused on curriculum standards and educator evaluation. These principles were based on the popular notion that teacher evaluation was wholly ineffective in identifying teachers who were not able to meaningfully increase student outcomes (Kane & Staiger, 2012). Influential organizations and partnerships such as the Measures of Effective Teaching Project (http://www.metproject.org/index.php) published a number of non-peer reviewed studies on teacher evaluation models that combine observation and value added metrics. Projects such as this suggested that the combination of value added scores and teacher observations increase the psychometric reliability and validity of the evaluation metric, allow for a close tie to student learning, and provide specific areas of teacher strength and weakness. Moreover, the combined system has been suggested to be far better than using constructs such as teacher experience or education level as proxies for quality teaching (Kane & Staiger, 2012). With these improvements in mind, the United States Department of Education (USDoE) specifically noted that the waiver’s purpose – through encouraging a shift to a new set of standards as well as educator evaluation - was to allow states and local education agencies (LEAs) the flexibility needed to improve instruction and student performance (United States Department of Education, 2012).

**Pennsylvania’s Teacher Effectiveness Model.** In Pennsylvania, school code and educational practice were adjusted in order to meet federal waiver requirements. Specifically,
the state adopted common academic standards and developed a standard evaluation tool for teachers, non-teaching educators, and principals. The latter, referred to as the Educator Effectiveness Models (EEM), were structured so as to include multiple measures of effectiveness based on direct observation of performance as well as student outcome data, including indicators of student growth using value added modeling (Pennsylvania Department of Education, 2014a). While educator evaluation had revolved around observations to some degree historically, the introduction of a standards-based observation model and the inclusion of student performance in teacher and principal evaluation marked a dramatic shift in both policy and practice.

**Reliability and validity of educator evaluation models.** The accountability models developed consequent to federal and state policy have been given a lot of attention by the research community. Much of this attention has focused on the psychometric properties of reliability and validity as they relate to the evaluative models directed at schools or individual educators. This focus is based on the idea that poorly constructed models are unlikely to lead to policy goals. Three of the four components of Pennsylvania’s EEM have extensive research related to reliability and validity. Concerns with teacher observation have focused on the consistency of ratings within a rater over multiple ratings and between raters who have observed the same teaching behavior; as well as the limited research base validating standards-based observations (Bell, Gitomer, McCaffrey, Hamre, Tianta, & Qi, 2012). Concerns with value added modeling revolved around the assumption that a non-experimental statistical methodology can be so comprehensively designed as to allow for causal conclusions linking educators to student academic gains. Many are unconvinced that models, even with their current statistical complexity, have been validated to the degree necessary to reach such conclusions, especially in regard to high stakes decisions related to teacher (Amrein-Beardsley, 2008; Darling-Hammond,
Amrein-Beardsley, & Rothstein, 2012), principal (Fuller & Hollingworth, 2014), or school performance. Similar to value added modeling, concerns with school evaluation metrics are widespread. Such metrics are traditionally based on available quantitative data, with the majority of weight typically assigned to indicators of student performance on high stakes tests (Hanushek & Raymond, 2005; Kane & Straiger, 2002; Ladd, 2001; Polikoff, McEachin, Wrabel, & Duque, 2014; Tobias, 2004). This limited focus raises concerns as to both the construct validity and generalizability of school evaluation (Linn, 2000). Some research efforts have looked at the consequences of accountability systems in use, noting that local responses can affect what is taught and how it is taught. In some cases, pressure on systems and individuals has driven a few to engage in dishonest practices in order to achieve the appearance of accountability driven outcomes (Hill, 2009; Jennings & Bearak, 2014). This line of research is of note in that it extends past the evaluation model itself and shifts attention to the role of the local actor in implementing accountability policy.

**Role of the central office in reform implementation.** While the technical properties of evaluation models are clearly important when it comes to understanding the likelihood of reforms leading to their stated goals, there are additional variables that should also be considered. A critical yet under-studied variable is the role a school district’s central office plays in facilitating implementation. Central offices “make decisions about numerous instructional guidance instruments including staff development, curriculum guides, curricular materials, teacher supervision, and student assessment” (Spillane, 2000; p. 142). Though limited attention has been given to district-level decisions about reform (Rorrer, Skrla, & Scheurich, 2008), some theoretical and empirical structure has been discovered. In general, this is based on the assumption that reform adoption is not a simple process of encoding and applying new
information; but rather, a complex cognitive process of learning, adapting, and revising that is heavily influenced by specific factors (as well as their interaction) within and surrounding a central office. Such an assumption suggests that while reliability and validity are important, alone their technical adequacy (or lack of) will not predict reform success. Rather, there are human variables that will act as confounds if not considered.

The specifics as to how staff in a central office learn, adapt, and revise new reform information influences implementation at the local level. First order (Honig, 2003), or form focused (Spillane, 2000) change occurs when the behaviors of an organization are adjusted somewhat to align with the new information, but the emphasis is on surface level change and appearance only – not a change in beliefs or organizational goals. Second order, or function focused change, occurs when the emphasis is on creating different and improved outcomes, and as such, the change is more aligned with the spirit of the original policy and reform. It is this type of implementation when paired with technically adequate models that is more likely to lead to policy goals. Unfortunately, form focused, or first order change has plagued education-based reform and is often linked to either a lack of capacity at the local level or a purposeful local resistance (Spillane, 2000).

Capacity can relate to the physical context, the individual, or the group (Spillane & Thompson, 1997). Physical capacity refers to having enough time, staff, and resources to understand and support a reform. Individual capacity deals with having staff that are intent on learning new ideas, have the background knowledge and experience to understand those ideas, and have the wherewithal to know best the way to present new information in the local context. Social capacity allows individuals to network within a larger group, reference interpretations against the local norms and culture, and shapes a common understanding across the district.
Resistance to change often manifests when capacity is limited (Ogawa, Sandholtz, Martinez-Flores, & Scribner, 2003; Olsen & Sexton, 2009), leading only to first order or form focused change.

**Theoretical Frame**

An appreciation of how accountability models are understood at the local level is vital for predicting the likelihood of such a model leading to second order change. However, the variables that affect this understanding, as well as the endless interactions between the variables, make conceptualizing such a process difficult. Fortunately, many of the variables that are introduced by central office staff – both as individuals and as groups - can be categorized within the frame of what is referred to as “sensemaking” in the literature (Coburn, 2001; Honig & Hatch, 2004; Seashore, Febey, & Schroeder, 2005; Spillane, 2000). Sensemaking is the process where interpretations of outside information are developed in a way that leads to local response (decisions, policies, etc.) (Seashore, Febey, & Schroeder, 2005). At the individual level, sensemaking, like learning, begins with exposure to and reflection upon new information. New information is paired with prior knowledge and initial sense, or understanding, is made. Individuals then share their sensemaking with each other and a group sense of what a reform means to the local group begins to take shape (Spillane, 2000).

The cognitive process of sensemaking has been given limited attention in the study of educational reform (Honig & Hatch, 2004; Rorrer, Skrla, & Scheurich, 2008). Research and theory from other fields has similarly focused on technical issues related to policy design instead of how those involved with change experience it; and how those experiences shape interpretation and sensemaking (Honig & Hatch, 2004). The narrow field of study surrounding sensemaking has paid the most attention to how background knowledge and social interaction influence
interpretation (Coburn, 2001; Coburn, 2005; Firestone & Shipps, 2005; Spillane, 2000; Spilllane, Reiser, and Gomez, 2006). Other factors such as the degree to which reform information is perceived to be important, useful, or lead to individual or group gain (Seashore-Lewis & Dentler, 1988), the influence of competing factors in a district (Firestone & Shipps, 2005), and the effect of personal beliefs of central office staff (Firestone & Shipps, 2005) have also been suggested to play a role in how a group understands and interprets new information. While these factors have been discovered in the noted research, there is very little understanding of how they (as well as any others) interact and, cumulatively, lead to the sensemaking of reform at a local level.

**Purpose of the Study & Research Question**

This study sought to extend the ongoing validation of teacher effectiveness models by examining the “how and why” of the central office experience and response to accountability policies. Such a perspective is understudied and not well understood; creating a gap in the knowledge of what, beyond psychometric reliability and validity, might predict the likelihood of accountability models being implemented in a way that will lead to their original intent. Studying this phenomenon provides additional information regarding the variables that impact implementation of accountability mandates. The more that is known about all the covariant factors influencing reform, the greater the likelihood that such policies will actually lead to their intended goal of greater student achievement. This study looked at specific local factors related to background knowledge, social interaction, perception of value or worth, competing issues, and personal beliefs; and asked how these factors influenced the sense that central office personnel in four cases made of the Teacher Effectiveness Model (TEM) mandate in Pennsylvania.
**Research Design**

The study used a qualitative, instrumental, multisite case study approach wherein data collection stemmed from interviews with central office administrators and review of documents. The core characteristics of qualitative methodology are described by Merriam (2009) as the focus on meaning and understanding (what it means to “know”), the researcher as the primary instrument, the use of an inductive process, and a focus on thick and rich description. All four of these characteristics aligned with the purpose of the current study through the attempt to understand the process of sensemaking at the central office level. Case study methodology allows for an in-depth description of a bounded system (Merriam, 2009). This study begged for in-depth description of central office staff responses to a particular reform mandate, which, while bound to the Teacher Effectiveness Model (TEM), could not be parsed away from the broader context in which central office staff live and work. This point should not be under-stated; indeed, it was the complexity of all the individual and context variables at the central office level that was the target of the current study. As such, a methodology that allowed for in-depth consideration of all that surrounds the central office staff was necessary.

Instrumental case study is appropriate when the focus is on the phenomenon being studied more than it is the specific case (Stake, 2005) and a multisite case study allows for a more compelling argument when considering findings (Merriam, 2009). This study was an instrumental multisite case study in that multiple central office responses were considered with the intent of learning how central office staff interpreted of accountability policy. The use of multiple cases allowed for similarities and differences across central offices to be identified and facilitated a more compelling conclusion.
In the current study, central office administrators across multiple districts were purposefully selected. The superintendent for each district was contacted by the researcher in regard to participation. As suggested by Bogdan and Biklen (in Merriam 2009), data collection occurred one site at a time. Thus, consequent to the first superintendent’s consent to participate, the researcher began with that district. Superintendents were asked to identify other key respondents, and interviewers continued until all had been approached. Upon conclusion of initial rounds of data collection, the next district was investigated. Review of documents related to the TEM also took place. Consistent with Yin (1981), Rist (1982), and Owens (1982), analysis of data were concurrent with collection.
Chapter 2: Literature Review

**Teacher Evaluation Background**

Historically, there have been a number of approaches to teacher evaluation. The specific methodology of such has been guided by philosophical beliefs as well as the desired outcomes, or purpose, of the evaluation. During the first half of the 20th century, the purpose of teacher evaluation was to ensure those leading students in classrooms were of solid moral character. Informal observation and interview were often the methods used to carry out the evaluation process (Ellett & Teddlie, 2003). As political dynamics in the country changed in the 1950s and 1960s, the purpose of teacher evaluation shifted the actual methods employed, with the assumption that the right instruction would lead to improved student performance. Behaviorism, as a psychological construct, was popular at the time, and likely influenced the development and wide spread use of formal observations (Darling-Hammond, Wise, & Pease, 1983; Ellett & Teddlie, 2003). The behavioral perspective, as a philosophical belief, assumes that the best ways to instruct students are independent of local context such as differences among schools, students, and/or communities. As such, supervisors can be trained to look for the “correct” ways to teach and teachers observed for how many of, or to what degree, these techniques are used. Despite the popularity of the behavioral perspective, other ideas on teacher evaluation were expressed during the second half of the 20th century. One concept rejected the idea that “good teaching” was defined by a set of universal traits, and rather, suggested that good teachers were able to reflect and respond to local conditions and specific student needs (Darling-Hammond, 1983). Another concept looked at the purpose of evaluation less as a means to judge performance and more as a way to identify strengths and weaknesses in a teacher’s approach, with results leading naturally to a continuous improvement process (Darling-Hammond, 1983).
Through the end of the 20th century, most decisions about purpose, philosophical belief, and, in turn, the specifics of teacher evaluation models were left to the local education agency. However, as education policy in general shifted with the passage of NCLB (2002), the goal of accountability overtook teacher evaluation (Hazi & Rucinski, 2005). Accountability, at face value, is less concerned with whether or not good teaching practice is universal or not, and rather, focuses heavily on student outcomes. Classes where students achieve at high levels or demonstrate adequate growth are assumed to have good teachers, and classes with poor student outcomes are assumed to have teachers in need of additional support (or more). The shift to accountability brought with it a focus on student performance on high stakes tests, and a statistical approach to gauge student growth labeled value added modeling (VAM).

Using student outcomes to evaluate teacher effectiveness has gained a great deal of support in the political and even philanthropic worlds; however, the psychometric criteria of reliability and validity have failed to be conclusively established (Steele, Hamilton, & Stecher, 2010). As a result, some have advocated for teacher evaluation models that combine teacher observation with student outcome measures. The shift away from using traditional observation checklists or observations of any type has been justified by a “growing consensus that teacher evaluation in the United States is fundamentally broken” (Kane & Staiger, 2012, p. 3). Traditional models have been found to under-identify struggling teachers and not efficiently differentiate between teachers who increase student results on tests and those who do not (Steele, Hamilton, & Stecher, 2010). The Measures of Effective Teaching project (http://www.metproject.org/index.php) has published a number of non-peer reviewed studies on teacher evaluation models that combine observation and value added models. Conclusions, which have proven to be very influential in federal education policy, have been made to suggest
that models with between one-third to one-half of an evaluation system rooted in student outcomes and the other portion based in a comprehensive observation system are the best way to identify effective teachers (Cantrell & Kane, 2013). Specifically, the combination of value added scores and teacher observations is suggested to increase the psychometric reliability and validity of the evaluation metric, allow for a close tie to student learning, and provide specific areas of teacher strength and weakness. Moreover, the combined system is thought to be far better than using constructs such as teacher experience or education level as proxies for quality teaching (Kane & Staiger, 2012).

**Background to Pennsylvania’s Teacher Effectiveness Model**

Accountability systems have a history at the state level dating back to the latter parts of the twentieth century; but were federalized through NCLB (2002) and the concept of Adequate Yearly Progress (AYP). Adequate Yearly Progress was a national attempt to ensure that all students were achieving at proficient levels. Unfortunately, and similar to earlier state attempts (McDermott, 2007), the mandates of NCLB (2002) including AYP were limited as it came to reaching their stated goals (Koretz, 2009; Jacob & Levitt, 2003; Johnson, 2011, Neal & Schanzenback, 2010). As such, the United States Department of Education (USDoE) (2012) issued an opportunity for state educational agencies (SEAs) to apply for a waiver to certain provisions of NCLB (2002). Thirteen provisions of NCLB (2002) were waiver eligible. Of particular interest were the provisions regarding the Adequate Yearly Progress (AYP) determination and timeline, as well as forced improvement requirements that resulted from AYP scores. In order to be considered for a waiver, SEAs had to provide a plan designed to meet four principles. The USDoE (2012) described the purpose of the four principles as to “increase the
quality of instruction for students and improve student academic achievement in the State and in LEAs” (p. 5).

The first principle addressed college and career ready expectations for students and required standards for English/Language Arts and Math, high quality assessments to measure achievement and growth in these standards, and a public reporting process of the impact of such standards. The second principle focused on state-level accountability and support for local educational agencies (LEAs). This principle required new Annual Measurable Objectives (AMOs) set by states for Title I schools to meet, public recognition of high and low performing Title I schools, evidence of closing the achievement gap, and support for schools through monitoring and technical assistance. The third principle addressed SEA support of effective instruction and leadership in schools, and required an evaluation system for school staff (teachers and principals) that used multiple measures to gage effectiveness; and was used to both guide improved instruction and to make personnel decisions. States were required to train everyone in teacher evaluation models and provide student growth data that would be part of the multiple measures. The fourth principle dealt with duplication and unnecessary burdens related to paperwork, and required states to “remove reporting requirements that have little or no impact on student outcomes” (p. 7). The waiver also required SEAs to consult with stakeholder groups in the state when developing the waiver application as well as to plan and encouraged SEAs to begin evaluating programs across schools to see what worked best in regard to the first three principles.

Pennsylvania applied for and was approved for a federal waiver in 2013. The Pennsylvania Department of Education [PDE] (n.d.) published a document outlining an overview of how the State would meet the principles necessary in the waiver. The PDE focused on three
of the four principles: college and career readiness, state driven recognition and accountability, and improving and supporting quality instruction and leadership. Each of these three principles was expanded from just Title I to all schools in the state. The approved, new, annual measurable objectives (AMOs) were 1) test participation rate, 2) graduation or attendance rate, 3) closing the achievement gap for all students, and 4) closing the achievement gap for historically underperforming students including students with disabilities, students from low socio-economic background, and students who were English Language Learners.

The State’s plan to reach the new AMOs included a new school report card and implementation of “Turnaround Principles” (PDE, n.d., p. 2). The report card, titled School Performance Profile (SPP), was a metric designed to allow stakeholders improved access and understanding of school performance. It replaced the AYP metric that had been used in response to initial NCLB (2002) requirements. Turnaround Principles included an educator effectiveness model for teachers and principals. These models were intended to be strong enough to lead to replacement of ineffective principals as needed, or provide principals with the flexibility to adjust scheduling, staff, curriculum, and budget. The models were also intended to allow schools to keep good teachers and prevent non-effective teachers from being moved from one school to another. Other Turnaround Principles in the State’s waiver application included adding time to the school day, week, or year; ensuring a research-based, rigorous, and standards aligned instructional program in schools; making data-based decisions; focusing on school safety including social, emotional, and health needs; as well as engaging communities and families. All of these strategies were to be implemented through the Pennsylvania Training and Technical Assistance Network (PaTTAN) and regional intermediate units (IUs) and supported through the State’s Standards-Aligned System (SAS) and Classroom Diagnostic Tests. The waiver also
noted Academic Recovery Liaisons (ARLs) who were to work with IUs and ensure Title I schools in need were getting support.

The core purpose of the federal waiver was to allow SEAs and LEAs more flexibility, which was to lead to greater student achievement outcomes. The PDE (2013b) sought to clarify how these outcomes would be measured. Pre-waiver measurement of student performance was based mainly on achievement data, and the evaluation of student performance was articulated through Adequate Yearly Progress (AYP) – a metric that looked at the total percentage of students scoring proficient or advanced on the Pennsylvania System of School Achievement (PSSA) tests as well as graduation, attendance, and test participation rates. Post-waiver measurement expanded to additional consideration of PDE’s new AMOs for Title I schools and from AYP to the School Performance Profile (SPP) for all schools. There were four designations for Title I schools, including Reward for high achievement and Reward for high progress or growth as well as Focus for the lowest 10% of all Title one schools and Priority for the lowest 5% of all Title I schools.

The SPP was described as expanding from AYP’s focus on achievement to also include student growth and closing the achievement gap. Achievement measures grew from just high stakes academic tests to also include industry-based tests, grade three reading performance, and SAT/ACT scores. Closing the achievement gap looked at all students in a school as well as historically underperforming groups, with the expectation that overall proficiency rates would increase over time (and thus, close the achievement gap). Student growth used a value added model to describe a schools “impact” (PDE, 2013, p 4) on student learning from year to year. The SPP added other indicators of performance such as graduation rate, promotion rate,
attendance, course rigor, and PSAT/Plan participation; and gave schools extra credit for the percentage of students scoring in the advanced range on certain achievement tests.

Annual measurable outcomes and much of the SPP relied heavily on student achievement data. Around the same time as the shift from pre to post waiver requirements, the PDE also revised its academic standards and the tools used to measure them (PDE, 2013a). Historically, the state had used the PA Academic Standards; however, a shift to other, more rigorous standards was consistent with federal waiver requirements for college and career ready standards. Pennsylvania’s response was to implement the PA Core Standards. As this shift to new standards was made, the assessments used also needed to be updated so as to ensure alignment between instruction and measurement. New standards and assessments were described by the PDE to be more rigorous and included eliminating a modified version of testing for some students with special education and a writing assessment during high school, as well as adding a writing component to grade three through eight English Language Arts. Also changed was a move away from PSSA tests at the high school level (considered a more general assessment) to Keystone assessments (considered course specific, and thus, more rigorous) in Algebra, Literacy, and Biology. The PDE used a two-stage plan of implementation to first shift to Keystone assessments at the high school level, and then PA Core Standards aligned PSSAs in grades three through eight (PDE, 2013c).

Components of Pennsylvania’s Educator Effectiveness Model

Pennsylvania’s School Code was originally made law in 1949 and has been amended on a number of occasions since. Section 1123 of the Code relates to the evaluation of educational staff, was most recently amended in 2012 (Pennsylvania House Bill No. 1901), and is most commonly referred to as Act 82. The most recent amendments of the Code were meaningful in
that the characteristics for which educators are evaluated were specified, the authority for development of the specific tool(s) to use for observation were explicitly given to the PDE and the introduction of a multi-measured component of student achievement was mandated.

The 2012 amendments detailed four areas in which classroom observation and instructional practice were to be evaluated for teaching staff, including planning and preparation, classroom environment, instruction, and professional responsibilities. Similarly, principal observation was guided by consideration of four areas, including planning and preparation, school environment, delivery of service, and professional development. These concepts were weighted at 50% for both professional staff and principal evaluation results. The additional 50% of a staff or principal rating was based on student performance. The addition represented a meaningful shift from prior rating systems. Multiple measures of student achievement including “performance on assessment, value-added assessment system made available by the department” (Pennsylvania House Bill No. 1901), and other PDE approved measures were listed as sources for student performance in the amended School Code. Further definition of assessment specified the “Pennsylvania System of School Assessment, the Keystone Exam, or an equivalent local assessment or another test established by the State Board of Education to meet the requirements of section 2603-B(d) (10) (i) and required under the No Child Left Behind Act of 2001 or its successor” (Pennsylvania House Bill No. 1901).

The task of developing the rating tool for teaching, non-teaching, and principal positions was mandated to the PDE with specific timelines for completion and implementation. This task included the observation system, as well as defining each of the observation terms and operationalizing the behaviors that should be observed to provide evidence of these areas. The PDE was also charged with defining the multiple measure component of the rating system
(Pennsylvania House Bill No. 1901). The PDE’s work as charged is outlined in the Educator Effectiveness Administrative Manual (PDE, 2014a). Three effectiveness models are detailed, including one for teaching professionals, one for non-teaching professionals, and one for principals. Specific criteria defining each of these categories are given as well as a number of forms and templates an LEA can use to complete an evaluation. In all three models, a final score is derived from sub-components and one of four evaluative labels is assigned – distinguished, proficient, needs improvement, or failing. From these four labels a final decision is made as to satisfactory or unsatisfactory performance.

**Observation.** All three evaluations give at least 50% of the overall evaluation score to observation. For teaching-professionals, the PDE adopted Charlotte Danielson’s Framework for Teaching model. The Danielson model was first published in 1996 and has undergone revisions in 2007, 2011, and 2013. The 2007 edition, among other things, expanded the use of the rubrics to specialists within the school including librarians, nurses, and school counselors. The 2011 revisions were done to allow the observation model to be included in the Gates Foundation’s Measures of Effective Teaching project. The most recent updates were a response to the increased rigor introduced by the Common Core State Standards (Danielson, 2013). In general, the model uses a four-point scaled rubric to guide evaluators over 22 components organized into four domains. These domains mirror, exactly, the four areas listed in Pennsylvania’s Act 82 – planning and preparation, classroom environment, instruction, and professional responsibilities. Danielson rubrics provide a great number of examples of behaviors for each of the four levels. The PDE, through its Standards Aligned System portal (www.pdesas.org), published guiding questions and examples for a number of non-typical roles so that the teacher version of the Danielson model could be used – a requirement of Act 82 (PDE, 2014a). These roles include
teachers of students receiving special education for autism, life skills, visual impairments, giftedness, learning and emotional disabilities, and speech and language delays; as well as teachers in career and technical education, English as a second language education, instructional coaches, and reading specialists.

Danielson’s manual (2013) describes how the model should be used for classroom teachers as well as the non-typical roles listed above. A typical observation event would begin with a pre-conference between the evaluator and targeted educator. The educator would go over their purpose for the lesson and the events anticipated necessary to reach that purpose. Using the planning and preparation domain as a guide, the evaluator would look for evidence of content and pedagogy knowledge, knowledge of students and the learning process, the ability to set student learning goals, knowledge of resources available, a coherently designed lesson, and a way to measure student learning. Following the pre-conference, the evaluator would observe the lesson being taught. Using the classroom environment and instruction domains, the evaluator would look for evidence of positive peer and adult interactions, a culture for learning, effective classroom and student management and productive use of physical space. The observer would also look for meaningful communication between the teacher and students, quality questioning and discussion, actively engaged students, the use of assessment to guide instruction, and the flexibility to respond to the many variables that can impact student learning. The observation event ends with a post conference wherein the evaluator reflects on the lesson with the educator, looking for evidence of reflective teaching, accurate records, communication with families, engagement with other professionals in the field (other teachers or experts outside of the school setting), professional growth, and a professional presentation. Following this conference, both the evaluator and educator consider the full four domain rubric and rate the educator on all 22 of
the components. Often, the evaluator and educator meet once again to discuss disagreements and come to a consensus (PDE, 2014a).

While Act 82 required a similar observation rubric approach for all educators, the traditional teacher model for Danielson could not be applied to all who work in schools and districts (PDE, 2014a). Consequently, the Department of Education developed similar types of rubrics for dental hygienists, school counselors, home and school visitors, instructional technology specialists, school nurses, and school psychologists; as well as for supervisory positions related to curriculum and instruction, pupil services, special education, and vocational education. Each of these rubrics is unique to the specific role; however, the general four domain structure and emphasis on evidence gathering and educator reflection on practice remain.

Building level administrators are given different status in Act 82, and thus, the Department of Education developed a separate Framework for Leadership. This is structurally similar to Danielson’s model but specifically notes strategic and cultural leadership, systems leadership, leadership for learning, and professional and community leadership. Unlike the structured pre-conference, observation, post-conference framework of the teacher model, evidence collection for building level administrators is not addressed by the Department of Education. None-the-less, evaluators of building administrators look for evidence of goal setting, data informed decision making, collaborate work environments, celebration of success, product resource management, adherence to policy and law, high expectations for students and staff, effective communication and conflict management, school safety, instructional leadership, community engagement, and professional growth.

**Student Outcome Data.** Both the teaching-professional and principal models give 50% of the evaluation to student outcome data, while the non-teaching professional assigns only 20%
(observation is 80% in the non-teaching model). Outcome data for teachers comes from three areas – building level, teacher-specific, and elective. Building level data is provided in its entirety from the school’s SPP score, and makes up 15% of the whole model. Teacher specific data, another 15% of the total model, comes from a variety of combinations of four sources, depending on the types of students a teacher instructs. These sources include performance on assessments, growth, progress through IEP goals, and locally developed rubrics. Growth data specifically references the value added model used in Pennsylvania - the Pennsylvania Value Added Assessment System (PVAAS), and comprises the majority of weight in the teacher specific data category when it can be applied (the educator must instruct in a “tested subject” in grades four through eight or a Keystone subject in high school). The manual notes that PVAAS data “should fairly represent the proportion of instructional responsibility that a teacher has for each student for a teacher for each state assessment” (p. 20). The LEA must determine the percent of instructional responsibility for each teacher; a two-part process based on the date of a student’s enrollment and time spent with a teacher on a tested subject. Elective data, 20% of the model, comes from locally developed measures of student achievement articulated through the Student Learning Objective (SLO) process.

Principal data also comes from three sources. Similar to teachers, 15% of the whole model comes from the SPP. Another 15% comes from what is labeled correlation data based on teacher-level measures. Such is further defined as a principal’s ability to understand the relationship between teacher level data and teacher observation, his/her ability to explain such a relationship, and an articulated plan to use data to support schools and the LEA. The final 20% of the principal model comes from a similar SLO process as is found in the teaching-professional model (Giles-Rudawski, 2014).
Pennsylvania’s School Performance Profile. Pennsylvania’s School Performance Profile (SPP) is based in part on guidance from the USDoE (2013). Federal guidance on state and local report cards mandate that all Title I schools must report student achievement data, test participation data, a comparison between student results and state AMOs, other academic indicators including graduation rate and those listed by the State in its waiver application, and teacher quality data. States were free to include additional information and data as long as it was “statistically reliable” (United States Department of Education, 2013, p 26). Report cards were mandated for distribution to stakeholders on an annual basis.

The SPP is described in depth on the State’s web page (www.education.state.pa.us). Content on the website describes the purpose of the SPP as well as provides details on how the overall metric is calculated. The purpose of the SPP is four fold, in that it is designed to 1) contribute to the State’s educator effectiveness system, 2) meet federal Title I accountability expectations, 3) inform the public of academic performance in each school in the state, and 4) be a resource “to communicate and compare performance, analyze performance indicators as related to achievement, and encourage best practice” (www.education.state.pa.us/).

The SPP is comprised of five indicators, 1) Academic Performance, 2) Closing the Achievement Gap for All Students, 3) Closing the Achievement Gap for Historically Underperforming Students, 4) Academic Growth, and 5) Other Academic Indicators. Academic Performance is based on the percent proficient or advanced on the PSSA or Keystone exam, the percent competent or advanced on industry standards-based competency assessments, the percent proficient or advanced in reading on 3rd grade PSSA, and SAT/ACT benchmark. Each of these areas is weighted with the total accounting for 40% of the overall SPP. There are different configurations of how each area is weighted based on the type of school. In all configurations
except comprehensive career and technical centers, PSSA or Keystone performance in math, reading, science and writing make up between 30% and 40% of the SPP (www.education.state.pa.us/).

Closing the Achievement Gap for all Students and for Historically Underperforming Students are worth 5% each in the total SPP. This metric is computed by taking baseline performance subtracted from 100%. The goal is to cut the total deficit between baseline performance and 100% in half over six years. Once baseline is subtracted from 100%, the result is divided by two, and then divided by six to end at the amount of growth needed each year for a six year period. Academic Growth is based on the PVAAS average growth index for a school and covers the same four academic areas as does Academic Performance. Each of these areas is weighted at 10%, and the growth indicator in general makes up 40% of the total SPP. Other Academic Indicators make up the final 10% of the SPP, and include cohort graduation or promotion rate, attendance rate, Advanced Placement, International Baccalaureate, or college credit, and PSAT/Plan participation. There is also an extra credit category where schools can add additional points to their SPP for the percent of students scoring advanced in the four academic areas or an industry test, and the percent scoring three or higher on AP exams (www.education.state.pa.us/).

**Pennsylvania’s Value Added Model.** The value added model used in Pennsylvania (PVAAS) was developed by the same group behind the Tennessee model (TVAAS) – SAS EVAAS. SAS EVAAS describes itself as providing, “analytical services, including value-added modeling and projection analysis, for the assessment of schooling effectiveness at the district, school, and, when requested, at the classroom level” (Wright, White, Sanders, & Rivers, 2010, p. 2). There are two types of value added models used in Pennsylvania. Both models are linear
mixed models and include variables for a student’s test history, known and estimated fixed effects, random and estimated random effects, and random variation.

The multivariate repeated measures (MRM) model is similar to a multivariate repeated-measures ANOVA and is used to estimate effects when there are multiple and consecutive years of data (Wright et al., 2010). Specifically, this model is used in grades four through eight on tests of reading and math - as each of these tests is given every year from third through eighth grade (PDE, 2014b). Because the tests used for input are not scaled the same way each year, results are converted to Normal Curve Equivalents (NCE). The NCE allows for a common scale on an equal interval by converting percentiles to z scores and z scores to NCEs (Wright et al., 2010).

The philosophy used in Pennsylvania suggests that students should, at the very least, not lose ground over time. This philosophy is operationalized by maintaining at least an equivalent NCE over multiple years. The previous year’s NCE is computed by taking into account all of the student’s prior test results (each weighted differently). The current year’s NCE is computed by taking the current and past performance to estimate a current level of achievement (PDE, 2014b). Students who earn NCEs less than the year before are considered to have made less than a year’s worth of growth (after taking standard error into consideration), and students who earn NCEs greater than the year before are considered to have made more than a year’s worth of growth (PDE, 2014b). Using the MRM model and this philosophy, the value added metric can be said to be based on student gains over the course of a year.

A univariate response model (URM), similar to an analysis of covariance, is used when tests are not given in consecutive grades. In Pennsylvania, science tests are given just once in elementary school and again only once in middle school. End of course assessments are given following Biology, Literature, and Algebra, most often during high school. The MRM model
will not fit these data, and thus, a “predictive methodology” (PDE, 2014b, p. 16) instead of a gain method must be used. The URM works by taking into account a student’s history of performance, considering other students from a larger pool with similar histories, and using that group’s performance on the current test, predicting what score a student should earn. Value added is computed by finding the difference between a student’s actual score and the predicted score (PDE, 2014b).

For both methods, teacher effect is reported as a function of standard error (SE). Growth that is +/- one SE is given a score of zero, where zero is equivalent to making the growth that was expected (a year’s worth). Growth that is more than one but less than two SE above zero provides “moderate evidence” that growth was more than expected. Growth that is two or more SE from zero provides “significant evidence” that growth was more than expected. The same guidelines are used for scores more than one but less than two SE below zero and two or more SE below zero (PDE, 2014b).

**Pennsylvania’s Student Learning Objectives.** The student learning objectives (SLO) process was adopted as a way to “document a measure of educator effectiveness based on student achievement of content standards” (PDE SLO Annotated Notes, 2014). Student learning objectives are expected to measure more directly from the content an educator teaches than do more general high stakes assessments and allow for measurement of student performance in classes where high stakes test scores and growth do not apply, such as grades K through second or high school courses outside of Algebra, English/Language Arts, and Biology (PDE SLO Annotated Notes, 2014). The Department of Education’s website ([www.education.state.pa.us](http://www.education.state.pa.us)) lists several example measures including district designed exams, nationally recognized standardized tests, industry certificate exams, student projects or portfolios. An extensive set of
resources available on the Departments Standards Aligned System portal (www.pdesas.org) reviews how to design, build, and review SLOs. A number of example templates are made available, as well.

The mandated template for SLOs has five sections. The first section documents demographics related to the educator and students. The second section records the actual SLO goal, the PA standard(s) to which it is linked, and a rational for why that SLO is important. Section three focuses on the measures used to indicate performance on the SLO goal. More than one measure can be used. For each, an indication of what the measure is, whether it will be used to gauge achievement mastery or growth, how often it will be administered, if there will be accommodations given to certain subgroups, how it will be scored, and how performance will be reported are all noted. The fourth section documents indicators of performance; specifically, the targeted achievement levels or amount of growth expected. The fifth section builds from the fourth to compute how many students met the performance indicator expectations and categorizes the result into one of four categories, similar to the other parts of the Effectiveness model.

An example of a math-based SLO for first grade is available (among others) on the Standards Aligned System’s webpage (www.pdesas.org). The goal statement relates to the mathematical relationships among numbers and how such can be represented, compared and communicated. In the example, the Everyday Math Assessment is listed as the performance measure and marked as a nationally recognized standardized test. The focus of the SLO is growth and the assessment will be given three times within the school year. The performance indicator is one year’s worth of growth as defined in the Everyday Math Assessment system. The final section of the form notes that at least 85% of students need to make the growth mark
for the teacher to earn a proficient rating, and 95% to earn a distinguished rating. Other SLO examples available on the Standards Aligned System’s webpage show the use of student projects and performance measured through rubrics or academic knowledge measured through locally developed assessments as likely sources of SLO data.

The SLO Users Guide (PDE, 2014b) states that performance measures should be “based upon high-quality performance measures aligned to the targeted content standards” (p. 14). Annotated notes indicate that measures need to be of “high quality” so as to know if students have mastered content standards (PDE SLO Annotated Notes, 2014). Principles of well-developed measures include alignment to content standards, standardization, rigor and fairness, objectivity, reliability, and validity.

**PA Model Summary.** Pennsylvania’s Teacher Effectiveness Model applies to almost all professionals working in schools or districts. Models are tailored to address the differences between teachers of different types of students, non-classroom teachers, and administrators; but all have the same basic four-part structure. An observation component makes up 50% of all the EEMs. The teacher observation is based on the nationally recognized Danielson Framework for Teaching and other observation models were developed by the PDE. For all, there is a standardized rubric with four domains and many subcomponents. All EMMs also take into account student outcome data, which is operationalized through school level performance, student growth, and student learning objectives.

Reviewing the details of the EMMs demonstrates the complexity of the educator evaluation approach. One cause for such complexity is the need to build models that accurately and consistently measure how effective the educators of Pennsylvania actually are. In order to remain aligned with the intent of federal waivers, which place student learning as a top priority,
decision makers must be sure that they are working with a reliable and valid evaluation of educator effectiveness. The following section considers these two constructs, reliability and validity, and the extent to which they have been manifest in similar models.

Reliability and Validity of Model Components

The Standards for Educational and Psychological Testing (American Education Research Association, American Psychological Association, & National Council on Measurement in Education, [AERA, APA, & NCME] 2014) are a set of industry standards intended to “promote sound testing practices and to provide a basis for evaluating the quality of those practices” (p. 1). The 2014 edition of the Standards take special care to address program evaluation and accountability metrics such as the Teacher Effectiveness Model; or components used within, such as school metrics like the SPP or growth metrics like VAM. Because of the high stakes involved with the outcome of such models and metrics for districts, schools, and individual educators, the Standards (2014) stress that, “it is essential that all tests used in accountability, program evaluation, or policy contexts meet the standards for validity, reliability, and fairness appropriate to the intended test score interpretations and use” (p. 207).

Kane and Staiger (2002) wrote, “In the rush to implement accountability systems, little attention has been paid to the imprecision of the test score measures on which they are based” (94). Their research, along with that of many others, has documented a number of technical concerns with accountability metrics. Indeed, Cizek (2009) noted that “large scale student achievement tests like those required by the No Child Left Behind Act (2001) typically have two purposes: (a) global estimations of an individual student’s competence in an area; and (b) accountability for student progress at an aggregated level” (p. 65). Cizek (2002) argued that as such, large scale tests, serving as information gathering procedures, should be evaluated using
the criteria of reliability and validity. It follows that as accountability metrics have replaced single tests as the ways in which schools are held accountable, the same criteria would follow (AERA, APA, NCME, 2014; Amrein-Beardsley, 2008; Fuller & Hollingworth, 2014; Hill, 2009; Hill, Kapitula, & Umland, 2011).

The most recent edition of the Standards for Educational and Psychological Testing [Standards] (AERA, APA, NCME, 2014) define reliability in terms of consistency. Consistent results are important so that any change over time can be attributed to a real change in the construct that is being measured, rather than some type of measurement error. The Standards (AERA, APA, NCME, 2014) define validity as “the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (p. 11). In other words, any test is designed to estimate skill set relative to some general construct and validity refers to how well the test actually executes this task (measures the construct). Validation, the act of collecting evidence to confirm or disconfirm the test’s ability to measure the construct, takes place over time and considers potential problems with construct underrepresentation (where the test does not measure all the important components of the construct) and construct-irrelevance (where the test measures variables outside of the construct). The Standards describe validity as a “unitary concept” (p. 14) and differentiate the evidence collected through the validation process. Other Committees, such as the National Research Council (2002), describe different types of validity such as content, criterion, construct, and consequence. The difference in definition between the committees may be somewhat arbitrary, as the Committee on Standards for Educational and Psychological Testing (APA, AERA, NCME, 2014) write about evidence on alignment (content), relationship with other measures of the same construct (criterion), measuring only the construct and not other, irrelevant, variables (construct) and consequences.
**Teacher Observation.** Historically, teacher observation has been the traditional method for teacher evaluation; however, with an increased focus on teacher quality and the link between teaching and student outcomes, concerns with the traditional observation approach have been raised (Kane & Staiger, 2012). New, more structured systems for observation, labeled “standards-based teacher evaluation” tools by some (Kimball & Milanowski, 2009, p 36), have gotten a lot of attention. These so called standards-based observation systems are thought to improve the psychometric properties of observation by establishing common expectations for teaching behavior based on professional literature and expert opinion (Kimball & Milanowski, 2009). As well, these models are designed to provide useful feedback that can be formatively applied at the teacher level for ongoing professional development and job performance improvement (Danielson, 2013).

From an accountability perspective, the reliability and validity of standards-based observation models has been questioned. Concerns with reliability have focused on the consistency of ratings within a rater over multiple ratings and between raters who have observed the same teaching behavior. Specifically, the confounding variable introduced by the relationship between the rater and the teacher could bias ratings, as could the attitude, or buy in, of the rater to the system as a whole. In addition, the rater’s skill in implementing the rating system could cause reliability problems. Finally, the influence of other teachers could bias an individual’s ratings – in that a school with many low performing teachers would artificially inflate the rating of average teachers – making them look much better than they actually are (Kimball & Milanowski, 2009). Limited research exists where these concepts have been studied, but some have suggested addressing reliability issues through thorough training for raters, using multiple ratings, and using multiple raters (Kane & Staiger, 2012). Others have suggested
extending the study of reliability from just inner-rater to examine the whole observation system, including the observation tool, the raters using the tool, the training for the raters, and how scores are actually made based on observed behaviors (Hill, Charalambous, & Kraft, 2012).

Bell and colleagues wrote of standards-based observation tools, “Unlike other fields that have used the same instruments for many years and systematically studied those instruments thoroughly across contexts and populations, the theory building and empirical aspects of the measurement of teaching under the now-dominant sociocultural, nonbehaviorist models of learning are nascent” (Bell, Gitomer, McCaffrey, Hamre, Tianta, & Qi, 2012, p 85). These authors then suggested a validity model for teacher observation models that took into consideration four factors. First, the evaluation of scoring should be considered in terms how appropriately the tool measures what it claims to measure (teacher quality), how accurately/consistently the tool is over multiple observations, observers, and time periods, and how biased or unbiased the tool and use of the tool prove to be. Second, Bell et al. (2012) noted how well the tool can be generalized as an important validity factor. In practice, observations are limited, and there is concern that a session spent observing math instruction might not fairly represent a teacher’s ability to teach reading or another subject. Third, the correlation between results on the observation tool and other evaluative forms of teacher/teaching quality such as value added modeling should have a positive relationship. Fourth, the implementation effects (what others have called consequence validity (AERA, APA, NCME, 2014; Hill, 2009)), or what actually comes from use of the tool should be considered. In the case of standards-based observation systems, which have both formative and summative usages, validity study should ask if teachers improve their performance in areas of weakness and are poor teachers removed from classrooms (Bell et al., 2012).
Empirical study of standards-based observations have focused most on Bell et al.’s (2012) third idea, the relationship between teacher ratings and student outcomes as measured by value added models. Kane and Staiger (2012), in a non-peer reviewed article, looked at this relationship across five different standards based observation systems including Danielson’s Framework for Teaching (FFT), Pianta’s Classroom Assessment Scoring System (CLASS), Grossman’s Protocol for Language Arts Teaching Observation (PLATO), Hill’s Mathematical Quality of Instruction (MQT), and Marder and Walkinton’s UTeach Teacher Observation Protocol (UTOP). The authors reported correlations between rating scales ranging from 0.67 (between the FFT and MQI) to 0.93 (between the FFT and PLATO). When looking at the relationship between student gain measured with value added modeling and observation, correlations ranged from 0.09 (using the MQI) to 0.27 (using the UTOP). Similar studies have found correlations between the Danielson model and VAM between 0.2 and 0.25 (Bell et al., 2012) and 0.27 to 0.43 (Milanowski, 2004), and between PLATO and VAM ranging from 0.09 to 0.16 (Grossman, Cohen, Ronfeldt, & Brown, 2014).

School Metrics. Traditionally, the focus of school accountability metrics, and thus, what is at the heart of comparisons, has been student performance (either level or growth, or both) on achievement tests (Hanushek & Raymond, 2005; Kane & Straiger, 2002; Ladd, 2001; Polikoff, McEachin, Wrabel, & Duque, 2014; Tobias, 2004). Hanushek (2005) cited a working paper by the same author that linked cognitive skills, standardized tests, and income differences, suggesting that standardized tests measured the cognitive skills students needed to compete in the marketplace. Such a link justified the emphasis on achievement tests, and motivated Hanushek (2005) to defend the models used in the mid 2000s by writing, “the skills observed to count in the marketplace are the ones forming the basis for school accountability” (p. 69).
Ladd (2001) described three different ways outcomes have typically been measured. The most basic approach is to measure a school’s average test score. In example, such would read as “School X had an average performance of Y on the test.” While this is a rather simplistic task, and thus, rather easy to implement, there is no control for influencing variables built into the model. As such, no reasonable conclusion linking one factor, such as how well a school teaches students, to actual student outcomes can be reached. The second approach considers the changing average score over time, with an emphasis on eventually meeting a specific target. If School X scored Y in year one, their goal by year five might be Y+5. This is an improvement over the first option in that it looks at performance over time, rather than just one snapshot; however, there are at least two meaningful problems. First, the growth target is often a benchmark imposed from the outside, independent of the actual school. As such, reaching that goal over time might not be logical given the resources of the school. Secondly, high performing schools may perform so high that there is a ceiling to their growth and low performing schools might have to make much more growth than would be reasonable for any school. The third approach is the Value Added Model (VAM), wherein a more complex statistical approach is used to show how much value the school actually contributed to student learning. Value Added Modeling is described in more depth below, but Ladd’s (2001) concerns with the approach deal with how “school” is defined, suggesting that VAMs may not be able to tease out the effects of teachers, school resources, peers, and the local community.

While these “other” variables, as noted by Ladd (2001) have been linked to school performance and/or confound school performance metrics, accountability systems continue to focus primarily on student achievement or growth on high stakes tests, even after federal waivers for flexibility on Elementary and Secondary Education Act (ESEA) allowed states to incorporate
additional indicators (Polikoff et al., 2014). Such an operational definition of outcomes in school accountability systems might be described by at least four reasons, 1) tests are relatively inexpensive compared to other types of reform efforts, 2) tests are easy to mandate in that the only observation need be of results, not process, 3) testing can be implemented quickly, which is politically advantageous, and 4) test results are easy to see and understand (Linn, 2000).

Much has been written about the potential problems with using high stakes tests as the cornerstone of school accountability. There are concerns with the assessments themselves - the content they measure and leave out (Linn, 2000). Ladd (2001) explains this point by writing, a “fundamental concern [that] arises in connection with these administered programs is their heavy reliance on student test scores, usually in just a few key subject areas, such as math and reading” (p. 385). There are concerns with what variables might also be considered in any accountability model – variables that might also contribute to school performance, and there are concerns with validity, impact, and credibility of an accountability system in general – do outcomes really reflect change in student learning or are they also explained by fluctuation in results, cheating, the exclusion of certain students, etc. (Linn, 2000).

Another concern with school metrics is the threat to consistency that random error can play. Random error can be attributed either to the subject taking the assessment or to factors outside of the subject. In example, things like motivation, attention, or full application of skill are related to the subject taking the assessment. Different testing conditions or different raters are related to outside-of-subject error. Such error, or inconsistency, can be measured either by the correlation between scores on the same test or between raters of the same subject, or by the Standard Error of the Measure (SEM). When correlations are low or the SEM is high, it is difficult to generalize results or have confidence that results accurately reflect skill related to the
The Standards (AERA, APA, NCME, 2014) note that the higher the stakes of the decision, the more precision necessary in the assessment(s) being used.

Kane and Staiger (2002) describe persistent and non-persistent causes to school outcomes. These two categories frame the sources of variance in a school accountability metric. Non-persistent variables, such as sampling variation or other acute factors fluctuate from year to year. The introduction of non-persistent variables results in a reliability concern as one cannot assume results would be the same year in and year out if no other factor (like instruction) were introduced. Persistent variables are relatively constant between years, and include both the intent of accountability evaluation, school performance in terms of curriculum and instruction, but also other long term factors such as community demographics. These “other” factors introduce error, or noise, into results and generate a (construct) validity concern.

The size of the school appears to play a confounding role in accountability metrics, introducing problems with too much variability from year to year – a reliability issue. Kane and Staiger (2002), specifically, considered the fluctuation, or variance, in accountability evaluations when those evaluations are based solely on student test performance. These authors reasoned that if accountability metrics really measured what they claimed to measure (student learning and performance) then results should be rather stable year after year. That is, the correlation between outcomes in one year should vary by only a little the following year. However, if the accountability metric includes a great deal of residual, or error, then the metric will fluctuate much more year to year, thus reducing the correlation between years. Using data from North Carolina, Kane and Staiger (2002) found that small schools showed the most fluctuation from year to year. Small schools were the most likely to be at both the top and bottom of the state evaluation system, from a rank order perspective; and large schools were
more likely to be somewhere in the middle. In North Carolina at the time, this meant that small schools were disproportionately more likely to receive either rewards or sanctions based on their results. Kane and Staiger’s (2002) research found that this was based more on sampling variation, as a result of the size of the school, than actual curriculum or instruction related variables. Confirming the problem, Tobias (2004) found the same type of issue in California while studying the 1999 Public Schools Accountability Act and resulting Academic Performance Index (API), a metric designed to evaluate California schools and drive an incentive program based on school improvement. The API was found to over reward low performing schools and under reward high performing schools. As well, the API fluctuated a great amount from one year to the next among small schools.

**Value Added Modeling.** Inherit to accountability is the need to link an outcome, or effect, (student learning) to a cause (teaching). In order to make a causal conclusion, that is, to be confident that one thing actually caused the other, all other alternative explanations that might also explain the outcome must be ruled out. In experimental designs, ample sample size, replication, randomization, and blocking are used to rule out all other potentials causes save for the one under study. Of these, randomization is held in high regard, as it is assumed to balance out all other potential confounding factors (Kutner, Nachtsheim, Neter, & Li, 2005). School and teacher accountability metrics are based on observed, not experimental, data; however, and using these types of approaches to rule out alternative explanations is not possible. As a result, as Kutner et al. (2005) write, “a cause-and-effect relationship between the explanatory factors and the outcome or response variable is difficult to establish in an observational study” (p. 645).

Despite such difficulty, value added modelling attempts to do just this, make a causal conclusion based on observed data. While there are different approaches to value added
modeling (VAM), the general approach is to account for alternative explanations by incorporating them into a complex regression model from the same family of models as Hierarchical Linear Modeling (HLM), multilevels, and random effects (Wiley, 2006). The purpose of the complexity is to deal with the nested structure of schools – that is, all the other variables that could influence student learning as well as the interactions of those variables with each other and the output. Variables that are most often added to the VAM include a student’s prior test results, prior teacher effects, and other student-level variables (American Statistical Association, 2014). The technical complexity of this idea should not be understated. Traditionally, social science has assumed that only through truly experimental designs could causal inferences be made. The VAM challenges this assumption by arguing that all other confounding variables can be accounted for in the model.

Wiley (2006) describes different types of VAM. The gain score model subtracts a student’s prior performance score from current performance to determine how much was learned over a period of time (usually a school year). This growth metric is computed for all students in a classroom and averaged. This average is then compared to the overall average growth for a particular set of teachers (in the school, district, or state). The difference between any particular teacher’s average and that of the overall average becomes the teacher effect. The Texas Growth Index (TGI) is an example of the gain score model in practice. A criticism of the gain score model is the built-in assumption that the effects of previous teachers persist over time. The covariate adjustment model (similar to the gain score model in that it looks at just two data points) addresses this criticism by allowing teacher effects to fade over time (by estimating previous teacher effects). The Dallas Value Added Assessment System (DVAAS) is an example of a covariate adjustment model. The DVAAS also takes into consideration student and school
characteristics such as race and SES. In this way, the covariate model attempts to make up for the lack of randomization.

A criticism of both the gain score and covariate adjustment models is the limited scope of data that is used (limited to just two data points). The layered model is much more complex in that it includes data over different subject areas and years, thus, accounting for both student performance over time and the layered effects of past teachers (Doran & Izumi, 2004). Because of the use of much more data, the most popular layered system, EVAAS by SAS Institute Inc., suggests that there is no need to adjust for school or student characteristics, and thus, the problems introduced by lack of randomization are addressed. The use of many data points also allows for estimating missing data, a solution not available in the gain score or covariate adjustment models. The Tennessee Value Added Assessment System (TVAAS) is the most famous layered model. It is notably different than the first two models in that it uses an average effect over three years to estimate teacher effect (Wiley, 2006).

A number of criticisms of the layered model have spurred the development of other, somewhat similar, value added models. Concern that the assumption of teacher effects remain consistent over time is an error has been addressed by the persistence model. This model has been designed to estimate a rate of teacher persistence. It also addresses concerns that a comprehensive history of test scores might not be able to control for student characteristics by making adjustments for such as well as student growth each year. The cumulative within-child model also responds to concerns that student variables play a confounding role that is not controlled through test history inclusion. While these two models – persistence and cumulative within-child – offer potential improvements on the traditional layered model, none were used in practice at the time of the Wiley (2006) article.
Statistical Assumptions. As noted, the suggested advantage to value added modeling is the ability to make up for a lack of randomization by way of accounting for alternative explanations for student growth (or lack thereof) beyond teacher or school effects. In order to make such a claim, VAMs must incorporate a number of assumptions. Reardon and Raudenbush (2008) argued that in order for causal conclusions as to teacher effect to be made, six statistical assumptions should be met. First, any student should be able to be assigned to any school in a comparison group. This idea links to random assignment and is vital for causal conclusion. The authors note that this is not the reality in the public school system, and as a result, comparisons between schools should be limited to schools with similar student characteristics. The second assumption is that the assignment of one student to a school should have no effect on any other student in the school, or that some students have no effect on other students. This idea links to confounding variables. The authors note that this is another assumption that is difficult to meet given that different students bring different backgrounds to schools and classrooms, and as such, this likely affects the degree to which certain topics are taught and/or the pace of instruction.

The third assumption is that test results are on an interval scale, a necessary characteristic in order to compare means. The authors note that there is no real way to know if learning is actually on an interval scale, as we are merely making inference to a cognitive skill and have no gold standard to which to compare. The fourth assumption is that of homogeneity, or the idea that a good school is good for every student, regardless of student specific characteristics. While this assumption is hard to make, the fifth assumption, labeled ignore-ability, can help meet the criteria for the fourth – wherein an X variable is added to the model to account for confounds. The idea is that all schools will have students with similar differences, and in the end, comparing across schools remains feasible. The authors note concern that identifying all the covariates is
difficult, and thus, the lack of random assignment probably catches up to the VAM through this assumption. The sixth assumption addresses extrapolation or the need for VAM models to predict performance beyond their data set. This is particularly relevant in schools with very little of a certain population, like a high or low free and reduced lunch population. Reardon and Raudenbush (2008) concluded that VAM are far from clearly meeting these assumptions; however, they note that VAM that incorporate meaningful covariates have the best chance of correctly attributing teacher effects.

Raudenbush (2004) listed six technical concerns with VAM, including how best to adjust for covariates, whether to treat school and teacher effects as fixed or random, how to represent the cumulative effects of teachers and schools, how to model covariation in student responses and teacher effects, how to incorporate multiple cohorts, and how to deal with missing data. Raudenbush (2004) surveyed the literature in the early part of the century and concluded that while VAM appeared to be a useful model in allowing conclusions regarding school effects, VAM had not progressed to a point where effects of the school context and specific teacher instruction could be differentiated from each other. Comparing between schools using VAM was appropriate when student characteristics in those schools were similar.

**Concerns.** A great deal of attention has been given to the Reardon and Raudenbush (2008) assumptions and Raudenbush’s (2004) technical concerns; as well as how well VAM can actually back up the claim of causal conclusions. The biggest assumption made in the VAM process relates to making a causal conclusion from observed data. Many are unconvinced that models, even with their current statistical complexity, have been validated to the degree necessary to reach such conclusions, especially in regard to high stakes decisions related to teacher (Amrein-Beardsley, 2008; Darling-Hammond, Amrein-Beardsley, & Rothstein, 2012),
principal (Fuller & Hollingworth, 2014), or school performance. Looking at EVAAS specifically, Amrein-Beardsley (2008) noted a lack of peer review supporting the methods used to account for missing data of the assumption that student test history could be used in place of accounting for student background. Moreover, Amrein-Beardsley (2008) expressed concern about claims made by the EVAAS model relating to teacher effect or future student performance (post-secondary), writing, “there is no evidence that research reports conducted internally, and especially externally, have validated such claims” (p. 66).

Others have noted specific psychometric concerns with VAM. In a policy brief commenting on the problems with using test scores to evaluate teachers, Baker et al. (2010) cite primary research findings regarding the consistency of teacher effect metrics from one year to the next, non-random assignment of students to teachers, and the limited generalizability of the actual tests used in VAM calculations. Additional concerns found in the literature include 1) questionable alignment, or content validity, resulting from the limited scope of achievement tests that become the fodder for VAMs (Grossman, 2014), 2) how well VAM outcomes correlate with other measures of teacher evaluation like observations (Bell et al., 2012; Grossman, Gohen, Ronfeldt, & Brown, 2014; Grossman, Loeb, & Wychoff, 2013; Strong, Gargani, & Hacifazlioğlu, 2011), instructional alignment to standards (Polikoff & Porter, 2014) or professional certifications (Harris & Sass, 2008), and 3) how well other variables that could become alternative explanations for student gain (or lack thereof) are addressed in VAM (Amrein-Beardsley, 2008; Darling-Hammond, Amrein-Beardsley, & Rothstein, 2012, Hill, Kapitula, & Umland, 2011). Looking specifically with VAM use in principal evaluation systems, Fuller and Hollingworth (2014) noted concerns with a lack of validation in using tests designed to measure student achievement for principal effect as well as with the structure of
models, which often fail to take into account factors such as prior student performance, student characteristics, school characteristics, and principal tenure.

While seemingly a minority in the academic context, there are some who believe the assumptions made are sound and that causal conclusions about teacher and school effect can in fact be legitimately made. Evidence from peer reviewed publications include Ballou, Sanders, and Wright’s (2004) test of the assumption in EVAAS models wherein accounting for student test history serves as a proxy for student background variables finding negligible differences between models that explicitly account for student background and those that do not (when the models are consistent with the general EVAAS structure with test history). The authors concluded that, as a result, teacher effectiveness conclusions are reasonable. Additional argument from the creators of EVAAS disputes criticisms relating to 1) limitations of the achievement tests that create the input for VAM, 2) missing student data cannot be validly addressed in VAM, 3) class size itself is a confounding variable that, like missing student data, cannot be validly addressed in VAM, 4) using student test history in place of student background variables does not fully account for the latter, 5) the EVAAS model is not transparent and/or is too complex for most to understand, 6) EVAAS methodology is not peer reviewed, and 7) EVAAS predictions of student performance have not been validated (Sanders, Wright, Rivers, Leandro, 2009). As noted by Amrein-Beardsley (2008), however, there is limited empirical support for these arguments.

**Student Learning Objective Research.** While there is ample evidence in the empirical literature based on the reliability and validity of measures similar to the School Performance Profile or Value Added model used in the Teacher Effectiveness Model, Student Learning Objectives remain somewhat elusive in terms of understanding the research base upon which
they are founded. General research on SLOs is extremely limited. Two peer-reviewed works report mixed results in terms of criterion-related validity. While Goldhaber & Walch (2012) found teachers with better performance on SLOs were somewhat more likely to also have better VAM outcomes, Balch & Springer (2014) found no significant relationship between the two.

**Consequence Validity.** There is debate as to the extent to which the consequences of tests or measures should be included in framing validity. The Standards (AERA, APA, & NCME, 2014) differentiate the improper use of a test result from unintended consequences unrelated to the test. The former, exemplified when a test validated to measure math proficiency is used to evaluate the effectiveness of math instruction, is considered a validity issue (measuring outside the purview of what the test was validated to measure). However, the latter, exemplified when a teacher or school/district narrows its math curriculum as a result of the math test used, is considered outside of the validity construct. Others have also noted the differences between these types of consequences (Moss, 1998); but none suggest that both types are not important in the overall evaluation of a test or measure.

Considering consequences, or validity in use (Fuller & Hollingworth, 2014) expands the focus of a measure, metric, or model from the factors within to the broader context where such is implemented (Moss, 1998). Shepard (as cited in Lane, 2014, p. 128) frames this as looking beyond the original intent of the measure and adding a focus on how the measure changes or impacts the broader system. Almost all research related to high stakes assessments and teacher evaluation models looks at the consequences of reform after it has been implemented. Lane (2014) commented on the consequences that occur after a test result is provided, listing factors ranging from changes in staff development, student engagement, use of results to adjust curriculum and instruction, and the impact of results on different subgroups within a school.
population. Others have found responses such as narrowing of the curriculum, coaching or explicitly teaching test-taking strategies (as opposed to core content), and actual cheating on tests (Hill, 2009; Jennings & Bearak, 2014). Examining teacher evaluation models with an emphasis on student outcomes, only limited and short term increases in student achievement have been found, with increased likelihood of teachers leaving less affluent schools and the gap between students relative to socio-economic status growing (Lavigne, 2014). While limited, study of models with both observation and student outcome data included has found that teachers report positive perspectives on the observation component (Jiang, Sporte, & Luppescu, 2015).

**Reliability and Validity Research Summary.** Cumulatively, the research on teacher evaluation models shows a great deal of work remains to be done when building models with technical adequacy. Standards based observation models have limited exposure in the empirical literature, but show concerns with consistency over time and between raters as well as with correlations to value added model metrics. School metrics have also shown problems with consistency and an over-reliance on a narrow set of inputs (high stakes tests), the latter leading to concerns as to the extent to which these measures actually comment on the overall performance of the school. Value added modeling is noted by many to not control for the plethora of variables influencing student growth to such an extent that cause-effect relationships between teachers and student learning can responsibly be made. Other concerns with narrow input (high stakes tests in just a few subject areas) as well as consistency have also been found. The student learning objective (SLO) construct, used in Pennsylvania’s teacher evaluation models, has an exceptionally limited research base, and at that, there are mixed results when correlated to other indicators of teacher performance within that limited base.
Moreover, research focused on consequence validity, or validity in use, shows that there likely are factors other than just traditional reliability and validity that impact how a model is implemented and the potential affect it might have on student outcomes. While some of these responses are positive and others clearly negative, the simple fact that tests and measures, including accountability systems, interact with and change the context in which they are implemented suggests that fully understanding any assessment must go beyond just technical reliability and validity research. Interaction with the local context introduces a need to take into account the human factor as a source of variance in predicting accountability model success, perhaps just as important as technical reliability and validity. Indeed, as Moss (1998) wrote in the context of evaluating the worth of measures, researchers should understand, “how individuals make sense of and use the products and practices of testing in their everyday lives” (p. 7). The following section moves in this direction, focusing on the individuals who often first interact with reform systems such as teacher effectiveness models.

**Role of Central Office Staff in Reform Implementation**

The journey of accountability-based reform from the state or national policy level to the classroom door makes an important pass through the school district’s central office. Often, it is teachers and schools who implement reform; however, teachers and schools are nested within their local school districts and thus, are influenced by a district’s unique context and characteristics (Seashore, Febey, & Schroeder, 2005; Spillane, 2000). District office staff “make decisions about numerous instructional guidance instruments including staff development, curriculum guides, curricular materials, teacher supervision, and student assessment” (Spillane, 2000; p. 142). While the research on district-level decisions about reform is somewhat limited
and discontinuous (Rorrer, Skrla, & Scheurich, 2008), some theoretical concepts have been proposed.

Reform does not come lock-stock from the policy level, but rather, is interpreted and fitted to the local context. Historically, a great deal of control and discretion has been afforded to the local level of administration. As such, while reform-based policies are written at the state or federal level, many details regarding implementation are made locally (Spillane, Reiser, & Reimer, 2002). Moreover, district level staff, unlike policy makers, have knowledge of the values, beliefs, and experiences within the district and are uniquely positioned to create an environment where reform principles might both be embraced and fully understood (Spillane & Thompson, 1997). The central office “decipher[s] what a policy means [in order] to decide whether and how to ignore, adapt, or adopt policy proposals into local policies and practices” (Spillane, 2000, p. 377). It is this local knowledge that allows district needs, strengths, and resources to be tailored for successful implementation (Honig, 2006), leading to development of the processes and structures and allocation of the human capacity needed (Sanders, 2012). Indeed, it is this translation and modification process that becomes the core role of the central office in the reform effort.

Honig (2003, 2006, 2008) proposed a two directional model for how reform oriented information is gathered and processed at the central office level. In the first direction, central office staff explores, scans, or is provided with information or evidence from outside of the district for the knowledge needed to understand a reform. This “search” for information is most effective when district staff are both able to understand the language and complexity of reform as well as filter the totality of the reform, understanding which specific parts of a reform might best fit within the district. The second direction looks inward, and covers how a district “uses” the
information gained from the search to shape reform development and implementation at the school level. Use can be considered at three levels – interpretation (making sense of information from the outside per a local perspective), storage and incorporation (making reform practices part of local policy), and retrieval (giving enough emphasis to reform behaviors that such are used to guide other decisions within the district). Interpretation can be seen through the discussions leaders have with each other about the reform, their buy-in to the goals of reform, and the attention and resources directed toward bringing the reform into the district. Incorporation is evidenced both formally through the development of rules, procedures, and policies that are rooted in new evidence and also informally through new behaviors and language used at the implementation level. Retrieval is marked by the shift from seeing evidence as new or different to something that is embedded in the culture and values of the system. Collectively, this process leads to the way in which any external reform is implemented at the local level.

First or Second Order Change. Due to the fact that reform implementation is dependent on how it is interpreted and used at the local level, it follows that actual implementation might go in different directions. A form-focus, or first order change, takes place when the day-to-day processes and procedures shift to something new but the overarching goals and desired outcomes of the organization remain the same. This can occur when district level staff attends to and begins to make sense of a reform. These understandings are at risk of missing the point of the reform. In a study of mathematics, Spillane (2000) found, “In attending mostly to the forms rather than the functions of the reforms, district leaders were making chiefly surface rather than structural analogs between the reforms and their experiences. Drawing surface analogs and failing to access the deeper structural relations between the reform proposals and their experiences, district leaders interpreted the reform in ways that missed its spirit” (p.
When the emphasis is placed on a function-focused, or a second order, change; the actual beliefs and values of the district change and the evaluation of reform is not the process itself but rather the original goal that spurred the reform (Honig, 2003, Spillane, 2000). In the former, the organization is driven by the goal of implementation; in the latter the organization is interested in creating a different outcome. Form-focused or first order change has plagued education-based reform and is often linked to either poorly designed policy or a purposeful local resistance, often driven by a lack of capacity (Spillane, 2000; Spillane, Reiser, & Gomez, 2006). Research on policy design is offered in the reliability and validity section, and an understanding of capacity and resistance is reviewed below.

**Capacity.** Spillane & Thompson (1997) detailed a model for local capacity as it relates to educational reform, dividing the concept into three types, human, social, and physical. Physical capacity refers to staffing, time, and materials. Time is a meaningful component of capacity when it comes to reform, as the learning process takes time to learn, implement, discuss, change, and evolve. Staff must have the time available to attend to new information, reflect upon it both individually and in group, and try the results of their sensemaking process in practice. Clearly, having enough staff so as to allow for individual time for such is important, as are the materials that accompany a reform.

Human capacity organizes the knowledge, commitment, and disposition of individuals within an organization. An intrinsic commitment at the individual level to the principles of reform leads to a motivation to learn the skills needed to develop and implement. Sanders (2012) echoed as much, suggesting that individuals must have the true intent to change as well as a deep understanding of both the reform and the change process. The process of sensemaking plays a vital role in individual capacity (Honig, 2008; Seashore, Febey, & Schroeder, 2005,
Seashore-Louis & Dentler, 1988; Spillane, 2000). When exposed to the overwhelming information that accompanies a reform, an individual must determine how much value any given piece of evidence should have and process the meaning of such within the local social, physical, and cultural context. This value judgment is influenced by an individual’s history and past experiences. When new information is related to old or familiar information, the individual is more likely to assign greater value than would happen with novice or unfamiliar information.

Effective social capacity leads to the spread of reform principles across the district, sustainable programming, and a shift of ownership from the central office to those at the building level (Coburn, 2001; Sanders, 2012). The spread of principles is related to Honig’s concept of use. The use of new knowledge through alignment with local values, incentive to change, and judgment as to the usefulness of the new information is most effective in a shared learning environment. Moreover, in the social setting, new ideas are refined and able to spread throughout the organization (Seashore-Louis & Dentler, 1988; Spillane et al., 2002). As Seashore-Louis and Dentler (1988) wrote, “individuals and groups must digest it, comprehend it and decide what to do with it” (p. 47). Studying the sensemaking process of teachers, Coburn (2001) found the conversations among peers, in both formal and informal settings, allows for the co-construction of understanding, collective decision-making, and negotiation on technical and practical points of implementation.

Sustainability, defined as the fidelity of maintenance of reform over time, is influenced by the depth of understanding of reform principles and the degree to which such understanding is embraced. This is heavily influenced by the social capacity of the district. To the same extent, the spread of reform principles and behavior is dependent on social networks within the district, the intent that drives implementation, and the depth to which principles are understood. Finally,
a shift in ownership to those implementing the reform takes place when new behaviors become a part of the everyday routine (Sanders, 2012).

**Resistance.** The individual, group, and physical capacity of a district affects the degree to which reforms are embraced or resisted. Districts with limited capacity are much more likely to resist change, and rather, put most of their limited resources toward form-focused implementation (Ogawa, Sandholtz, Martinez-Flores, & Scribner, 2003; Olsen & Sexton, 2009). Threat rigidity was described by Staw, Sandelands, and Dutton (1981) as a theoretical construct for explaining how organizations resist functional, or second order, change. In general, when a threat is introduced, an organization becomes rigid in its response by reverting to the behaviors it knows best. Staw et al. (1981) organized threat response into two categories, information processing and control processes. From an information processing perspective, the rigid response is marked by a restriction on information pertaining to ideas other than those widely embraced already. As a result, old behaviors are reinforced but new or alternative ideas are rarely considered. Often, there is only one approach to problem solving, there is a simplification of what is going on, there is a limit on cognitive complexity, and any search for additional information is limited to the ideas already being used. Staw et al. (1981) described control processes as being marked by increased centralization of authority, formalization, standardization of procedures, and greater control of decision making. The focus of the organization shifts toward efficiency and accountability and tolerance for creative or novel ideas is discouraged.

**Role of Central Office Summary.** As suggested by the literature relating to consequence validity, there are factors outside of accountability models that play into how successful those models will be when implemented. One source of this additional variance is the
human factor involved when reform moves from state or federal policy to the local context. Individuals and groups in school district central offices seek out information and put into use their own interpretations. As interpretations can be different from one local context to the next, implementation can generally be categorized as either a first order, form-focused change – where the emphasis is on surface level change only; or second order, function focused change – where the emphasis is on deep shifts in values and beliefs in order to fully embrace the spirit of reform. Which direction the implantation goes is affected by the district’s capacity. The following section details a theoretical frame through which the actual process of interpretation can be understood.

**Sensemaking as a Theoretical Frame**

The process of sensemaking provides a theoretical frame under which the many variables (and their interactions) associated with central office staff interpretation can be organized. When exposed to the overwhelming information that accompanies a reform, an individual must determine how much value any given piece of evidence should have and process the meaning of such within the local social, physical, and cultural context. Spillane (2000) framed this individual process as a cognitive constructivist activity and argued that such a theoretical frame allows a greater understanding of how policy transforms as it penetrates the local district context compared to a behaviorist frame wherein only stimulus and response (but not consideration of local sensemaking) is made. The cognitive frame, argued Spillane (2000) allows for a window into how leaders understand reform through their own learning process.

Consistent with Spillane’s cognitive constructivist concept, Honig and Hatch (2004) framed sensemaking as a process of crafting coherence. In such, individuals and groups within the local organization consume new information from the outside and set goals and strategies for
within. Part of this process is to simplify the new information so as to not overwhelm the organization and to link new to already existing practice. Goals and strategies are productive when they are constructed in a social, or collective, manner. A second component to crafting coherence is bridging and buffering new information. Bridging new to old is accomplished when those involved in the search process have relationships with outside experts and have the capacity to understand new jargon and link it to local language. Buffering relates to limiting how much new information gets through to the local level. Similar to concepts described in threat rigidity, buffering can lead to symbolic adoption of reform wherein some behaviors are changed but the general goals of the local organization remain the same.

Sensemaking can be thought of as the process where interpretations of outside information are processed in a way that leads to local response (decisions, policies, etc.) (Seashore, Febey, & Schroeder, 2005). In essence, this is part of the learning process and has both individual and group components. At the individual level, sensemaking, like learning, begins with exposure and reflection upon new information. New information is paired with prior knowledge and initial sense is made. Individuals then share their sensemaking with each other and a group sense of what a reform means to the local group begins to take shape. The group, or social, aspect of sensemaking has been found to play a key role in finding a common interpretation through both formal and informal group interactions (Seashore, Febey, & Schroeder, 2005).

At its core, sensemaking is the process of learning. Crossan, Lane, and White (1999) provide a framework that outlines what this learning process looks like starting at the individual level, shifting to the group, and eventually evolving the practice of the institution – just as Honig (2006) described the process of interpretation storage/incorporation, and retrieval. In the
Crossan et al. (1999) model, the concept of strategic renewal explains the tension between new information and that which the individual or group knows well. Strategic renewal embraces two key components, exploration and exploitation. Exploration, similar to Honig’s search, describes how an individual or group looks for new ways of doing things or ways to improve. Exploitation, similar to Honig’s retrieval, is the application of long standing practices often found in organizational routines and policies. Exploration begins with the individual intuiting new information - thinking about and making sense of what is presented. This sensemaking is the same as Seashore et al., (2005), Spillane (2000), and others have detailed. In the Crossan et al. (1999) model, individual intuiting or sensemaking leads to interpretation – just as Honig has also explained. Interpretation takes place at both an individual and group level. In essence, this is a continuation of the sensemaking process wherein individuals share their perspectives with the group, and in turn, the group begins to develop a common sense of the new idea or reform. Crossan et al.’s (1999) third stage considers the integration, or application, of new ideas into practice. This is a group activity where individuals try out new behaviors, reflect with the group and fine tune to fit their local context. Over time, the new behaviors are fitted to the organization and become embedded into the day-to-day practice. Successful integration leads to institutionalizing, wherein the whole system implements and regularly uses the once new ideas. At this point, ideas are not considered new, but rather, become part of the everyday culture and values of the system.

The cognitive process of sensemaking has been given limited attention in the study of educational reform (Honig & Hatch, 2004; Rorrer, Skrla, & Scheurich, 2008). Similarly, research and theory from other fields of study have focused on technical issues related to policy design (much like the focus on psycho-metric reliability and validity in school evaluation
research) instead of how those involved with change experience such, and how those experiences shape interpretation and sensemaking (Honig & Hatch, 2004). Some studies regarding the sensemaking process do exist, however, and when considered cumulatively offer greater insight on the complexity and importance of local interpretation when looking at reform implementation and potential success. Spillane, Reiser, and Gomez (2006) categorized this research into two broad areas, that related to existing background knowledge and that related to social processes.

**Background Knowledge.** Background knowledge can be considered at three different levels, all of which embrace the assumption that any new information an individual learns is considered relative to what the individual already knows. In a summative chapter on the topic, Spillane, Reiser, and Gomez (2006) consider the idea that some portion of variance related to why reforms fail in schools could be due to the following background knowledge-related ideas, as opposed to just poor reform design or active local resistance. These authors write about mental schemas, assimilation and accommodation, and the depth of knowledge an individual brings to reform learning.

A schema is a mental guide for understanding information and it is within this mental guide the new information is considered (Spillane, Reiser & Gomez, 2006). In example, a principal might have a pre-existing schema for what “good” classroom instruction should look like. When that principal is exposed to a new observation model, the rubrics and expectations of the model are not learned in isolation, but relative to the schema the principal already had in mind. Because individuals have different schemas for the same thing, based on belief systems, history and experience, etc., any new piece of information can be interpreted very differently simply based on the individual schemas with which the new information is framed. In review of a number of studies on standards implementation (see Cohen & Barnes, 1993; Jennings, 1992;
Spillane, 1996; Spillane, 1998; Spillane, Reiser, and Gomez (2006) write, “Teachers’ beliefs about subject matter, teaching, students, and learning were influential in what they interpreted from state and national standards about their practice” (p. 52). These authors note that these differences do not necessarily reflect active (or passive) resistance to a model, but rather, are sincerely guided by the schema with which the individual brings to the new information.

It is difficult to change a schema once it is developed. Spillane, Reiser, and Gomez (2006) detail the differences between assimilation and accommodation; noting that most often when new information is presented the individual remains with the basic schema from before, and only assimilates new details around existing information. Accommodation, on the other hand, is the development of different schemas based on new information. When learning about reform, the more complex and transformative the new information the greater the need to separate from old schemas and accommodate new. Spillane, Reiser and Gomez (2006) point out that this is not the natural cognitive process that most are inclined to take. Honig (2003) corroborates this concern and links successful assimilation or accommodation to individual capacity, noting that administrators must have the ability to understand when and how to fit new information into the already existing local system. Unfortunately, study of such a phenomena in teachers has found that more familiar concepts are assimilated with pre-existing schemas and more foreign concepts are actually just given less attention (Honig & Hatch’s, 2004) concept of buffering) during the interpretation process (Cohen, 1990).

As detailed above in the section on first or second order change, the amount of experience and knowledge an individual has with a reform the more likely implementation will go beyond a surface level attempt (Honig, 2003; Spillane, 2000; Spillane, Reiser, & Gomez, 2006). When individuals do not understand the complexity of a reform process or the greater
purpose behind those processes, implementation is driven by the goal of looking the part and simply ensuring the correct behavior. Looking at building level administrators and reading reform, Coburn (2005) found that the amount (and philosophical type) of reading background a principal had influenced how reading reform was interpreted. As a result, the type and amount of information about the reform that made it from building level administrators to teachers implementing the reform was influenced by the building leader’s background. Studying math reform, Spillane and Thompson (1997) found that not all administrators actually understood the state-level reforms that were introduced. While the administrators studied reporting being comfortable with the reform information, evidence suggested that the understanding was surface level, or process related, only and lead to implementation that was “superficial topical alignment” (p. 188). Similar conclusions were also found by Sanders (2012), further supporting the importance of a leader’s understanding of reform principles. Also related to background knowledge is personal belief; the norms, values, and expectations of individuals and the district as well as the local culture influence how sense is made (Firestone & Shipps, 2005).

Social Processes. The social process related to learning extends to the “organizational structure, professional affiliations, social networks, and traditions” (Spillane, Reiser, & Gomez, 2006, p. 56). This broad definition exemplifies the complex nesting of education where teachers work in classrooms, which are in buildings, which are in districts, which are in local communities, and so on. Spillane, Reiser, and Gomez (2006) argue that the social influence on interpreting new information is influenced at each of these levels as well as by interaction between the levels.

The quality of social interactions is related to the sensemaking process. Working with others can allow greater in-depth processing of the reform/change concepts and more expertise
and resource than simple self-reflection. Groups can make collective decisions as to which policies to follow and which to ignore and with the former are able to make the transition from abstract policy concept to concrete classroom-level behaviors. Through practice and discussion, groups allow for new practices to emerge and become part of the regular culture of the district (Coburn, 2001). Group work does not automatically lead to policy aligned sensemaking; however. When conversations are not connected to practice, when there is not enough time for in-depth conversations, and when the worldviews of the group are too different from reform ideas, new policy is rarely adopted as a second order change (Coburn, 2001; Seashore, 2005).

As well, factors specific to the building level administrator are important, including his or her social network with colleagues as well as how well the administrator balances his or her role as the mediator between the central office and classroom teachers (Spillane et al., 2002).

**Other.** Additional factors related to how sense is made of new information include the perception of usefulness and possible gain (value, worth, or benefit) of new information as well as the capacity of the district to implement. Each of these has meaningful overlap with factors noted above – background knowledge and social process. Seashore-Louis and Dentler (1988) framed the sensemaking process across three ideas (salience, use, and gain) and two levels (individual and group); 1) salience (Did individuals receive the new information in a way that left an impression that such was important?), 2) personal use (Did the new information help the individual?), 3) organizational use (Did the new information help the organization through new policies or practice?), 4 personal gain (Did the new information help the individual improve?), and 5) organizational gain (Did the new information help the organization improve?). Results of an analysis of the sensemaking process in the educational setting found four areas of note.
First, there should be a meaningful incentive to learn in order to encourage the learning process. At the individual level, learning can be motivated by performance incentives (a desire to do a better job), influence or authority incentives (sense of feeling more important) or material rewards (job security or career advancement). At the organizational level, learning can be motivated by state or federal mandates, pressure of other like organizations, a general need to solve problems or improve, rewards for the organization, and past experience with reform implementation processes. Spillane, Diamond, Hallett, Loyiso, Zoltners (2002) support this concept, finding that organizational incentives such as those based on test results influenced how sense was made of certain accountability mandates. Second, the degree to which new information is discussed and analyzed at the group level is meaningful. Such relates to Crossan et al.’s interpretation level and influences how fully new ideas can be linked to local values and adjusted to fit the local context. This social process takes time when done well, and thus, inter-relates to Spillane & Thompson (1997) concept of physical capacity. Third, the degree to which new information is usable in the local setting influences sensemaking. This is a perception based finding wherein the local perception of reform contributes to the value assigned, and thus, the attention and resource given to implementation. Fourth, the communication of new information is important. This starts at Honig’s search level where information from the outside must be clearly communicated to local staff. Communication continues as local staff translates the information to teachers at the classroom level. Seashore-Louis and Dentler (1988) concluded that the interaction between all of these variables contributed to implementation success.

Firestone and Shipps (2005) conceptualize district office sensemaking by considering multiple sources of accountability. This framework complements Seashore-Louis and Dentler (1998) by expanding upon their ideas of motivation to implement, match to local values, use-
ability, and communication relative to the political, bureaucratic, market, professional, and moral accountabilities that confront district leadership. All of these pressures must be given due weight when interpreting new information, and often it is only the district level administration who have the full picture of the totality of demands across all five of these areas. Similar to Honig and Hatch (2004), Firestone and Shipps (2005) note that such a diversity of demands creates great incoherence among accountability demands. To add to the complexity, the process of search is expanded to any actor(s) who might bring in new information, either formally, as many conceptual models suggest, or informally, as the authors noted happen when unions, school board or community members, or others bring their own concerns or interests to the attention of the central office. In the case of the latter, goals might be in direct contradiction to the formal reform efforts mandated at the state or federal level.

Cumulatively, sensemaking in the educational reform context appears to be influenced most by the background knowledge individuals or groups bring to a new idea as well as the quality of social networks involved with interpreting the new information. Other factors, such as how useful and beneficial the reform might be for a particular context as well as what else is going on in the district or local community appear to also play a role. Most research on the sensemaking process has looked at the depth of background knowledge an individual has (as opposed to cognitive areas like schemas and assimilation/accommodation) or how the social process of interpretation takes place. While some have written about the interaction between multiple factors (Rogoff, 1995, Spillane, 2000), formal study of such does not yet appear to be prevalent in the literature base.
Literature Review Summary

A review of the literature finds Pennsylvania’s Teacher Effectiveness Model to be structured under the shadow of federal NCLB (2002) waiver requirements. These models – differentiated for classroom teachers, specialists and non-classroom educators, as well as building level administrators, include components of a standards based observation, school-wide achievement metric, student growth as measured by value added modeling, and a process labeled student learning objectives. Research on these types of components focuses mainly on the reliability and validity inherit to the tests, measures, or metrics themselves. A number of concerns with each are noted and with little doubt a great deal of work need still be done to achieve technical adequacy. Additional study of accountability models in practice raises concerns with how such models interact with other variables in the local context. Labeled by some as consequence validity, the interaction of reform policy with the human factor has the potential to be just as important in predicting second order change as the traditional technical study of reliability and validity.

Most research on consequence validity focuses mainly on how the results of accountability measures change what is happening at the local level. While the cognitive perspective and sensemaking theory provide a framework to understand how sense might be made of new, reform-related, information, very little empirical investigation of how this actually works with individuals at the central office level exists. This gap is unfortunate, for as researchers and policy makers continue to identify and focus on improving the technical adequacy of accountability models, the potential covariate factor of local interpretation remains largely ignored. The next chapter details a methodology designed to address this gap.
Chapter 3: Methodology

Qualitative Rationale

While many have described the worth of quantitative research (Gage, 1989), there has been for some time a recognition that the traditional – or rationalistic (Owens, 1982) - approach is unable to effectively study certain aspects within education, including some of the most pressing questions related to fluidity, change, and process (Rist, 1982). There is consensus on the core characteristics of qualitative methodology and general agreement that such an approach offers a different and complimentary perspective to the rationalistic. These characteristics are described by Merriam (2009) as the focus on meaning and understanding, the researcher as the primary instrument, the use of an inductive process, and a focus on thick and rich description. All four of these characteristics complement the research question that stems from the purpose of the current study, and direct connections are detailed in the following sections.

Meaning and Understanding. Rist (1982) contrasted the qualitative approach to quantitative by defining what it means “to know” in terms of “experience, empathy, and involvement” (p. 440). Similarly, Owens (1982) described a naturalistic approach to inquiry as an alternative to the traditional quantitative in how the former comes to understand and “know” (p. 2) that which is being studied. Rist (1982) asked the question, “What is going on here?” (p. 440) and suggested a more process-focused inquiry by seeking to “describe and analyze in an ecologically valid manner the values, behaviors, settings and interactions of participants in educational settings” (p. 440). Considering Owen’s (1982) point that quantitative approaches can only pursue that which can be quantified and measured, Rist’s (1982) values, behaviors, settings, and interactions demonstrate a need for a more naturalistic approach.
Inherit to this naturalistic approach is the study of a phenomenon within the broader context in which it takes place, indeed, this study within the setting approach is in stark contrast to the traditional approach of isolating a limited number of variables for investigation. Owens (1982) wrote, “Human behavior is so significantly influenced by the context in which it occurs that regularities in those contexts are often more powerful in shaping behavior than differences among the individuals present” (p. 5). Thus, any study of a district-based phenomenon would necessitate a study within the district setting, as the phenomenon and the context are too “interrelated” to be separated. Owens (1982) goes as far as suggesting that the separation of that which is being studied from the context will actually lead to distorted findings. Likewise, Rist (1982) used the term “holistic” to describe looking at the whole as greater than the sum of its parts, or to think that understanding the interaction between just two variables within a larger system is not enough.

The focus on meaning and understanding in a qualitative perspective can also be described as interpretive (Merriam, 2009). As opposed to a rationalistic frame, wherein reality might be spelled with a capital $R$, the interpretivist standpoint looks to understand the reality as it is perceived by those under study, and assuming that different individuals will see things differently (reality spelled with many different lower case $r$’s). Merriam (2009) described this goal by writing, “The overall purpose is to understand how people make sense of their lives and their experiences” (23). Owens (1982) defined this idea as the “qualitative-phenomenological” hypothesis, writing, “Understanding behavior hinges on understanding how participants know and see themselves, their feelings, values, perceptions, and actions” (p 5).

The qualitative definition of meaning and understanding as a means to “describe and analyze in an ecologically valid manner the values, behaviors, settings and interactions of
participants in educational settings” (Rist 1982, p. 440) best fits the purpose of the current study as the purpose was “to know” how central office staff responded to accountability policy. This line of inquiry could not be answered simply by computing outcome data or studying numbers, but rather, had to be investigated holistically and descriptively. Such a pursuit required study within the context of the districts themselves, an approach that allowed data collection to reflect the lived experiences of those being studied and take into account “how participants know and see themselves, their feelings, values, perceptions, and actions” (Owens, 1982, p. 5)

**Researcher as Primary Instrument.** Merriam (2009) notes “the researcher is the primary instrument for data collection and analysis” (p. 7) in qualitative research. The idea of researcher as the instrument is in contrast to techniques typically used in quantitative methods such as surveys or tests. Merriam (2009) describes why this difference is important, noting that it allows for ongoing adaptation during data collection as well as for a fuller and deeper understanding of what is being studied as it is understood by those being studied. Both of Merriam’s points are described in more detail below (adaptation in Inductive Process and fuller and deeper understanding in Thick and Rick Description), but some general qualities of this characteristic are necessary to note. Merriam (2009) suggests the lack of structure that is inherit when the researcher is the primary tool for data collection and analysis. Unlike the traditional rationalistic methodology, such is seen as a great advantage in qualitative study as it allows for the flexibility to “adapt to unforeseen events and change direction in pursuit of meaning” (p 20). This ability to adapt is key to qualitative investigation; however, this does leave the world of structure, linearity, and predictability behind.

Flexibility was necessary for the current study. Data were collected on site through a discovery orientation, meaning that as data were collected and analyzed additional issues arose
and begged further inquiry. Given the complexity of studying response to policy within the broader context, it was impossible to predict prior to data collection everything that would need to be investigated. The researcher needed the flexibility of limited structure to allow data collection to actively and recursively guide additional collection.

**Inductive Process.** Rist (1982) described inductive logic as studying the specific and building to the general; this logic assumes to suspend judgment on what is important, critical, or relevant (as opposed to quantitative methods that define such in the problem of study). Rist (1982) suggested that even the problem at hand is hard to define initially, and rather, must be “discovered through the process of study” (p. 442). Merriam (2009) described an inductive strategy as building “toward theory based on data collection from the field” (p. 7) wherein findings could be grouped into themes or categories. Inherit to the inductive process is the flexibly allowed by using the researcher as the primary tool for data collection and analysis and the “emergent” (Merriam, 2009, p 8) design that naturally follows. The current study required such an inductive process in that it was unknown how central office staff would respond to accountability policies. As the perspectives of those who work in the central office were fundamental to the purpose of the study, knowing what was important, critical, or even relevant could only be discovered through immersion in the context.

**Thick and Rich Description.** Merriam (2009) described a thick and rich description as the outcome of data collection and analysis and as something that is presented in “words and pictures rather than numbers” (p. 8). Owens (1982) wrote that such a full and deep description, “provides meaning of human behavior in the real world in such terms as cultural norms, deep seated values and motives arising from cherished tradition, and community values” (p. 7). Moreover, Owens noted that such a presentation of findings “conveys a literal description that
figuratively transports the readers into the situation with a sense of insight, understanding, and illumination not only of the facts or the events in the case, but also the texture, the quality, and the power of the context as the participants in the situation experienced it” (p. 8). Such a full and deep presentation of data was vital in the current study. The literature base on sensemaking suggested that response was influenced by a number of variables, related both to the individual and the group, and a potential for complex interactions between these variables. This complexity went well beyond that which might be quantified and framed in a statistical equation; and rather, necessitated a think and rich qualitative description.

Case Study Rationale

Merriam, Stake, and Yin have published a number of works describing case study methodology. Merriam (2009) defined case study “as in-depth description and analysis of a bounded system” (p. 40) and further defines bounded system as the “unit around which there are boundaries” (p. 40). Stake (2006) defined “unit” in case study inquiry through the term “unit of analysis,” describing specifically as a particular program, classroom, etc. that is selected for study because of its typicality or uniqueness. Merriam (2009) expands the concept to include a process, issue, or concern. The concept of unit of study or analysis is important in case study as it is one of the core characteristics that serve to differentiate the method from other qualitative approaches. Indeed, Merriam (2009) writes that without a bounded unit, a study cannot be a case study.

Case study methodology is not defined, however, by the tools that are used to collect data, but rather, by the questions that are being asked (Yin, 2008). Stake (as described in Merriam, 2009) lists four characteristics of research questions that lead to effective case case study, including 1) a desire for information that is concrete rather than abstract, 2) a desire for
information that is rooted in the context in question, 3) a search for information can be further interpreted by the reader, and 4) a goal for the reader is able to apply what is found in the case study to his or her own familiar populations. Yin (2008) labels these “how” and “why” questions and further argues that the case study approach plays well when what is being studied cannot be separated from the context in which it is found or when the researcher has a limited amount of influence over the interplay between the unit of study and the broader context.

Merriam (2009) describes three special features of case study - particularistic, descriptive, and heuristic. Particularistic refers to the focus on the specific case at hand and how each unit is important because of the way in which the case informs the researcher about the phenomenon that is being studied. Descriptive refers to the concept of how a thick and rich description (see above) can explain a complex situation wherein a great number of variables influence the phenomenon over a long period of time. This form of description is how the researcher dealt with Yin’s (2008) caution of the difficulty of separating the unit of study from the broader context. Heuristic refers to how a specific (particularistic) case illuminates the unit of study as an example that can be used outside of the specific context to learn.

Merriam (2009) describes a number of different types of case study including historical and observational, intrinsic, instrumental, and collective, evaluative, and multisite. Instrumental case study is described by Stake (1995, 2005) as used when the focus is on the phenomenon being studied more than it is the specific case. The case itself may be a critical example in that in-depth analysis of the particular is designed to gain insight toward something more general. Collective and multisite case studies share similarities in their nature and description (Baxter and Jack, 2008). Merriam (2009) summaries multisite case studies as a method that looks across multiple cases in order to build a more “compelling” (p. 49) set of findings. Stake (2006) defines
multisite study as a number of individual cases with similar or common characteristics or conditions. Collectively, these cases should share the same unit of study and allow for deeper understanding of a particular phenomenon. Yin (2003) describes multiple-case study as a way to explore a phenomenon between cases and suggests that cases be purposively picked so as to find similar results across the different cases.

The unit of analysis that worked to “bound” the current study was the sensemaking of accountability policy (the Teacher Effectiveness Model) by central office administrators in school districts in Pennsylvania. Consistent with Merriam’s (2009) construct of “unit,” this was conceptualized as a process that had taken place over the period of time when accountability policies were first mandated until the time of data collection. The research questions were aligned with Stake’s (1981) types of questions in that the knowledge that was to be gained was concrete in nature, the knowledge that was to be gained was rooted in the context, and the knowledge that was gained was presented in a thick and rich description that allowed for further interpretation by the reader and generalized to reader-specific settings.

Consistent with Yin (2008), the sensemaking response to accountability policies was a process that could not be separated from the local context of the study due to the complexity of central office staff sensemaking and interpretation; indeed, it was the latter that provided the fodder for the specific focus of the study. This complexity could only be described in what Merriam (2009) labeled “descriptive,” in that a thick and rich description was necessary to fully explain the sensemaking characteristics and how these responses were influenced by the individual, local context, and interaction between each. As well, the current study was particularistic, with a specific focus on central office staff in Pennsylvania; however, the insight gained from studying the phenomenon of response to accountability will provide greater
knowledge as to the application of such theories in other, similar, settings. Finally, the cases being studied served to illuminate response to accountability policy as an example that can be used outside of the specific context.

As Merriam (2009) described, there are a number of different types of case study. Stake (1995, 2005) review instrumental case study as useful when the focus is more on the phenomenon being studied than the specific case itself. Stake (2006) and Yin (2003) discuss studying multiple cases in order to examine common characteristics and develop a deeper understanding of the phenomenon under study, and Merriam (2009) add the advantage of building a more compelling case by incorporating multiple sites. The current study was an instrumental multisite case study in that multiple central office responses were considered with the intent of learning how response characteristics to accountability policies could be framed. Each case was analyzed individually (a within case design) and a cross-case analysis looking at common themes across central office staff followed completion of the final case.

Site and Sample Selection

Merriam (2009) used the terms particularistic and heuristic to describe two special features of the case study. These terms can also be useful in defining the sample selection. In the current study, a multi-site study of districts in Pennsylvania allowed for specific focus on the similarities and differences of response to accountability policy across multiple local contexts. Such allowed for a broad understanding of response to accountability policy. The sample for the current study was built according to the following process. All districts in the four county region surrounding the Harrisburg area were listed and compared across total enrollment, percent economically disadvantaged, geographic size, pupil to teacher ratio, and locale. Any district that did not have a consistent superintendent over the bounded time frame of the study was
eliminated from the potential pool of districts. Working in districts that had not dealt with a superintendent transition allowed a clearer picture into what played into response and interpretation of accountability factors. From the remaining list of districts, a sample that crossed the listed factors was made to diversify the conditions under which sensemaking was studied.

Central office administrators from South Central Pennsylvania were purposefully selected for the current study. The superintendent for each selected district was contacted by the researcher in regard to participation. As suggested by Bogdan and Biklen (in Merriam 2009), data collection occurred one site at a time. Thus, consequent to the first superintendent’s consent to participate, the researcher begin with that district. Upon conclusion of initial rounds of data collection, the next district was investigated. Within each district, and while interviewing the superintendent, inquiry as to other key respondents was made and additional interviews followed as appropriate.

**Research Strategies**

Once sample districts were selected, the superintendents of each were contacted by phone or email. Initial contact attempted to establish common ground as educators and leaders, and move toward explaining the study and seeking initial interest in participating. A follow up email provided a more in-depth explanation of the purpose of the study and requirements of any district studied. Once district consent was gained, background information on the district was investigated through review of publically available documents and archives. Interviews with the superintendent were scheduled and as described above, other key respondents were identified and interviewed. The interview protocol is provided in Appendix A. All interviews were recorded and handwritten notes were made. Following an interview, the researcher listened to
the recording, ensured notes taken during the interview were accurate, and buttressed notes with additional information. A typed summary of the interview was made with additional thoughts and responses as well as answers to research questions included (analysis). The completion of a case was evident when all key respondents had been interviewed and additional information failed to provide new or useful information. A case study database (Yin, 2014) was used to keep all information collected including hand written notes, audio recordings, typed notes, thoughts, and reactions; as well as typed responses to research questions.

Building the trusting relationship necessary to obtain a deep level of information is a difficult process. It would be unfair to insinuate that any district would be analyzed only in a positive way; however, an emphasis on the purpose of the study, considering response to policy, was thought to be meaningful to leaders struggling to meet state-imposed mandates with ever increasing fiscal limitations. Moreover, in-depth study focused on understanding response as a lived-experience was in stark contrast to the objective data based measurement techniques currently used to evaluate districts. This contrast was welcomed by superintendents and served as added motivation to allow access. Rist (1982) also discusses the importance of establishing a trust or rapport with those being studied so as to gain the maximum amount of information possible. Initial conversations and questions that focused the study on response to policy from the perspective of the local lens went a long way in forging such a relationship.

Data Collection & Analysis

Collection and analysis of data in the naturalistic paradigm were co-occurring and dependent on each other. Yin (1981) wrote, “the investigator is also doing analysis during the data collection process” (p. 105); Rist (1982) wrote, “analysis occurs concurrent with as well as subsequent to data collection (p. 445); and Owens (1982) wrote, “data collection and analysis go
on simultaneously, with the analysis giving direction to the data collection by suggesting what to check, when to seek confirmation, and how to extend the data collection itself” (p. 11).

The three established methods for data collection in case study are interview, observation, and document analysis (Merriam, 2009; Rist, 1982; Yin, 1981). The research purpose of the current study suggested a reliance on interview, with additional information coming from document analysis. Rist (1982) reviewed characteristics of the interview in a qualitative pursuit. While interviews can be formal or informal, structured or semi structured, Rist (1982) suggested that they present more as conversations between the researcher and participant, with flexibility enough to adjust the direction, questions, and probes based on the information that is being shared. Document analysis, like the interview, should be open-ended and with limited constraint or exclusionary criteria. Again, the purpose is to explore the information that is available and find out as much as possible. Rist (1982) wrote that there “exist no rules that either include or exclude the use of any written sources” (p. 444), and lists any number of options ranging from formal records such as school files, tax records, public hearings, newspaper articles to more informal records such as interoffice memos, minutes from meetings, and personal notes (that are available).

Analysis in qualitative study is not sequential, or linear, and Rist (1982) was clear in noting such, writing, “the sequential approach is the Achilles heel of qualitative work. To pose the question, ‘Now that I have the data, what should I do with it?’ is not only the wrong question but erroneously defined qualitative inquiry” (p. 445). Rist (1982) pointed to two considerations, that analysis should be concurrent with data collection and that there is no single way to analyze what is discovered. In regard to the first concept, the researcher is constantly learning from the
data collected, so if an interviewee provides a certain response, the researcher “analyzes” what is said and as a result asks a more probing question.

While considering and controlling for personal biases, the researcher must allow the data collection process to direct the path of continued inquiry. Such an approach necessitates ongoing analysis during the collection process, engaging in a continuous cycle of collection, analysis, collection, and so on. This multi-directional approach is aimed to ensure a deep understanding of what is going on in the setting and with participants, and ensure that proscribed procedures do not get in the way of inquiry.

The collection of data comes from many sources, different interview questions that go to the same topic, interviews with different stakeholders, and various documents of relevance. Analysis looks to identify common themes across data collection. Rist (1982) described the interaction between sources using the term triangulation, which simply suggested comparing findings in one form (observation perhaps) to those in other forms (interview, document review). Triangulation adds to the general credibility of qualitative research by aiming to ensure that findings are representative of what is really going on. This is accomplished by looking for themes and trends within the data, similar reports from multiple respondents and evidence in observation consistent with data from interviews and document analysis.

Data collection and analysis for the current study was fit into a framework that linked the research question to theoretical factors of sensemaking, questions that drove the purpose of data collection, and possible evidence (Yin, 2014). The framework is found in Appendix B. Document analysis included external and public documents that provided a more comprehensive understanding of the variables district decision makers would have had to consider (including budget documents, school board meeting minutes, local media reports, and the like). Initial
interviews were scheduled with superintendents. These interviews were semi-structured, using
the questions detailed in the interview protocol (Appendix A) to remain focused on the research
topic, but allowing for flexibility in regard to additional probing questions. Analysis was
ongoing and information was framed according to the identified factors of sensemaking. An
isolated analysis of each case was made and a cross case analysis considering similarities and
differences between districts followed the final case.

Reliability and Validity

In the qualitative approach, reliability and validity are defined as the credibility and
trustworthiness, respectively, of the data collected. Owens (1982) documented six procedures
necessary in order to ensure credibility in qualitative study. First, the researcher should spend a
prolonged amount of time gathering data so as to progress to a point of deep understanding of
what is going on and to move past personal bias. Second, as detailed above, data should be
triangulated so that themes discovered are confirmed through multiple sources. Third, the data
collected from one stakeholder should be checked against that collected from another (defined by
Owens as Member Checks). Fourth, referential adequacy materials should be collected to help
the researcher understand the context in which observations are made and recall all that was
observed. Fifth, a thick description of what is found should be offered. Sixth, peer consultation
should be used as a “check” for the researcher and opportunity to step away from the setting and
run ideas past others who are “experts” in the naturalistic field to ensure the researcher is on the
right track.

Additional steps to ensure reliable and valid reporting have been offered. Yin (1981)
added another validity check by suggesting that the case study be reviewed by the significant
stakeholders and informants to ensure the facts that are reported truly represent the intent of their
experiences. Owens (1982) discussed how judgments and conclusions need to be logically related to the findings of data collection, with clear linkages from one to the other. Guba and Lincoln (1981) described an audit trail as keeping and organizing all the documents collected or generated through the study to ensure credibility to any outsider as well as allow for replication of the study if applicable.

The current study took steps to ensure trustworthiness and credibility through a number of methods. First, two pilot interviews were conducted with central office staff not involved in the actual study. The purpose of the pilot interviews was to clarify questions and develop an understanding for what kind of information the proposed questions would elicit (Weiss, 1994). Second, the researcher immersed himself in the context of the study through the interview, document review, and data analysis process. This immersion is evident in the rich and thick description found in the analysis. To ensure that the characteristics identified in one form of collection were representative of the reality of a district’s response to policy, confirmation with other interview questions, other interviewees, and within district level documents was sought. A post analysis review of findings by the superintendents and other staff was conducted to make sure facts were accurately reported. As well, review of the case study by university level experts in qualitative methods was conducted, consistent with Owen’s (1982) peer consultation.

**Limitations and the Role of the Researcher**

While case study is the ideal method for certain research questions, such as the current study that has sought to investigate a certain phenomenon (response to accountability policy), certain limitations should be noted. The greatest drawback to the research design is the lack of generalizability. Findings go a long way in describing that which has been studied but are not able to predict what might occur in other, even in similar, situations (Merriam, 2009). In
addition, case study methodology is ripe for issues that stem from the naturalistic approach. Data collection and analysis techniques are put squarely on the shoulders of the researcher, yet training in how to observe, interview, and construct findings in a report is not standard (Merriam, 2009). Thus, the researcher risks providing only a partial account instead of truly being holistic, and has a meaningful chance of exaggerating or over simplifying a situation. Ethics are also a concern that stems from the case study approach. As the researcher is the tool for collection and analysis, personal biases must be considered and taken into account as part of the process (Merriam, 2009). As the phenomenon or context under study is often politically sensitive, the researcher must be mindful of data sources and the repercussions of exposing certain sources by name. The current study addressed these and other potential limitations by allowing other, more experienced case study researchers to read and comment on the work. In addition, certain respondents had the opportunity to review the findings to ensure their intent was captured.

The specific role the researcher played as the tool for data collection and analysis must be recognized. In the current study, the researcher had worked in public education in Pennsylvania for eleven years, seven as a school psychologist and four as a central office director of student services and special education. The researcher lived through the sensemaking process of the Teacher Effectiveness Model in his current district as it was processed by the superintendent and assistant superintendent. Working closely with these individuals as they were exposed to new information, and as they discussed it among themselves, with colleagues in other districts, and with building level administrators created a preexisting context that in one way or another biased data collection and analysis in the proposed study.
Chapter Four: Results

District A

**Background.** All four school districts included in the study were located in south central Pennsylvania. District A was in a large suburban local (National Center for Education Statistics, 2015). The district was 16 square miles in size, with 3,800 hundred students attending eight schools (SPP). Academic achievement scores, as measured by the school performance profile (SPP), ranged from 76.6 to 90.8 across the schools. Twenty-nine percent of the student population was classified as economically disadvantaged. The annual budget for District A was around 55 million dollars and data from the state Department of Education suggested there was around 18 million dollars in fund balance (PDE, 2015). The district website listed 16 central office administrators and 12 building administrators.

A review of school board meeting minutes from the 2014/2015 found very little dissent among members or any topics that appeared controversial. All but one vote over the entire year was unanimous. Only a few public comments were made at board meetings. One community member expressed concern with a raise for the superintendent and another with an academic issue relating to the speaker’s daughter. A review of local media reports found a few articles related to a growing need for more classroom space - a need brought about by an ever increasing student enrollment.

The Superintendent and Assistant Superintendent for academics were interviewed for the study. The superintendent had been in his position for five years. Prior, he served as the assistant superintendent and also the high school principal in the district. He had experience in other districts as a science teacher, assistant principal, and building principal. In total, he had worked 23 years in education. The assistant superintendent had been in her position for five
years, three as a commissioned officer and two as an assistant to the superintendent. Prior, she had worked as a supervisor for elementary curriculum, as an instructional coach, and as an elementary teacher. In total, she had worked 20 years in education, all in District A.

**Perceptions of Value or Worth.** Evidence related to the perception of value- or worth-factors of importance, usefulness, and gain described by Seashore-Louis and Dentler (1998) were mixed in District A as it applied to the teacher effectiveness model (TEM). While the overall topic of teacher effectiveness was important to the district, the TEM did not live up to its original purpose in actually addressing under-performing teachers. Ironically, the model actually seemed to make it more difficult to address under-performance. However, the observation component of the TEM embraced important values in the district related to teacher feedback, reflection, and growth. As well, the observation portion of the model was aligned, at least in spots, with other district initiatives, and proved to be somewhat useful in moving those additional concepts forward.

Prior to TEM, the district used a supervision and evaluation model that was perceived from within to be very effective in dealing with under-performing teachers. The superintendent described the process as “incredibly effective” in how it was able to identify concerns and lead to a plan for growth. Part of the old system was the Performance Assistance and Remediation (PAR) team - a group of administrators and a teacher mentor who worked with a struggling teacher. The core focus of the team and its work was growth and improvement, but the structure also allowed a way to work with a teacher if it became clear that another assignment or even another career was the best option. The superintendent commented on the benefit of this structure, noting the improvements often witnessed in the classroom as well as with other teachers, who were able to see the support a struggling teacher would get. He reported that
typically three to five teachers a year would go through the PAR process, and prior to TEM when PAR was working very well, more teachers were addressed in a three year span than had been addressed in the 20 years prior. He described the system as setting emotions aside and looking at real performance issues in the classroom. While the PAR team and process was very time consuming, the outcomes were clear and aligned with the purpose of addressing under-performing teachers. The Assistant Superintendent described the old model as “manageable, easy to implement,” and noted that, “everyone understood it and liked it.”

The TEM did not provide any advantage related to working with under-performing teachers, and in fact, was thought to actually make this process more difficult. The superintendent expressed great frustration over this point. He noted that the original purpose of the TEM was to decrease the total percentage of teachers in the state rated as satisfactory, citing a comment made by a policy-maker around the time of Act 82 passage. He described this perspective as “missing the point” and being rather out of touch with the operations of school districts. Prior to the TEM, very few teachers in district A had been dismissed. The Superintendent recalled only two cases where he would not allow teachers to resign due to the severity of certain behaviors. Instead, when teachers simply were not cut out to do the job, the PAR team would work with them, allowing the teacher to realize that resignation and movement to another profession was in their best interest. The Superintendent noted that these were not bad people, they were not hurting kids physically or emotionally, they simply were not meant to teach. These cases, however, were not formally documented anywhere, and policy-makers only analyzing the percentage of satisfactory or unsatisfactory ratings would completely miss what was actually going on.
The addition of student outcome data to the evaluation process was originally intended to make dismissal easier and tie results directly to teacher effectiveness; but in practice, this made dismissal based on the TEM almost impossible. District A was a relatively high performing district and the SPP scores for all its schools, when incorporated into the TEM metric, kept overall ratings too high to justify an unsatisfactory rating. Other data-based parts of the TEM required a great deal of time and district resource, but provided no observed value-added. As an example, in what the superintendent described as “maddening bureaucratic data,” principals and teachers had to sit and figure out the extent to which students made growth on their IEP goals. This was a very time intensive task, yet only contributed an overall one percent to the whole of the evaluation model.

In what the Superintendent described as “ironic,” the district actually had to find ways to go around the TEM when dealing with under-performing teachers. In one recent example, a non-tenured teacher was struggling in the classroom. Under the TEM, it would have taken three ratings – or 18 months – to document unsatisfactory enough performance to move to dismissal. While the observation rubric was helpful in regard to conversations about areas of growth, the district went outside the TEM and used progressive discipline based on incompetence to move toward a more timely change. Using this approach was effective; however, the Superintendent expressed frustration over how the tool originally designed to address issues like this actually proved in practice to not only be no better than prior models, but actually to provide less support to the district.

The weakness of the previous model in District A dealt with how it struggled to allow non under-performing teachers to reflect, receive feedback, and ultimately grow as professionals. The Assistant Superintendent described the old model as holistic; looking at instruction in
general, but not specifically analyzing what was going on, and identifying components that might be the focus for continuous improvement. To this end, the TEM was viewed by the central office as important (growth was in important concept, see section on Personal Beliefs, below), useful (the old system did not do a great job for the majority of teachers) and beneficial (observations over two years of the TEM led the central office to think that most teachers were able to reflect on their practice and grow as a result).

The Assistant Superintendent noted two specific strengths with the observation side of the model. First, she described the Danielson rubric as “analytic” instead of “holistic” (as the old supervision model had been). The analysis of instruction allowed for feedback tied to specific components, which led to more meaningful feedback from principals and allowed for more effective reflection from teachers. Second, she described the Danielson rubric as research based, referencing the general acceptance of Danielson domains and components across the field of education, not just locally within District A. In the prior model, what was thought to be important was a local decision, often made by the observing principal. The Assistant Superintendent thought there was a place for this, but just as much thought the “outside” lens of Danielson was helpful too in that it allowed teachers to compare themselves to more general expectations.

The Danielson rubric also aligned with other initiatives in District A, making the gains of using it all the more meaningful. A major focus in the district was work on an aligned curriculum and assessment system (see section on Competing Factors). One of the components in the Danielson rubric focused on assessment. As such, one of the building principals was able to stress assessment during supervision meetings and worked to increase the awareness, reflection, and ultimately, practice related to assessment. The prior, more holistic, observation
model would not have allowed for such a seamless marriage between supervision and district initiative. Another focus in the district was a shift to the more rigorous PA Core Standards. Part of the standards’ rigor included increased student engagement and responsibility for their own learning, a concept also embraced in the Danielson model. Specifically, many of the differentiating markers between the “distinguished” and “proficient” ratings across the domains had to do with the degree to which students took responsibility for their own learning. As with the assessment example, the overlap allowed the district to increase the attention to and reflection upon a district initiative through the observation system.

The observation side of the model was not without its drawbacks, however. Central office staff had heard from some building principals, especially those with larger buildings and more staff to evaluate, that the time necessary to hold pre-conferences, conduct formal observations, document evidence observed, and hold post-conferences was exceptionally consuming. Given the amount of staff for which this process had to be repeated, it was hard for building administrators to make the process meaningful, which led to perceptions from some that there really was no more advantage in the new model compared to the old. This drawback was one of logistics more than it was the potential of the actual model.

On the eve of TEM, District A was not looking for nor thought they were in need of an improved model for supervision and evaluation. The concept of supervision, especially as it focused on growth, and evaluation, especially as it dealt with under-performing teachers, was addressed with the model that was already in place. As such, while these concepts were important to District A, there was no sense of a need for a new model to address it. Through early implementation, the District found no usefulness or gain in the model relative to working with or dismissing under-performing teachers and actually had to work outside of the model.
when such concerns arose. The increased opportunity for reflection, feedback, and growth that were associated with the observation component of the TEM were considered important ideas related to supervision, and showed early signs of usefulness and gain. Moreover, alignment between domains or components within the TEM and other district initiatives increased the perception of potential value or worth of at least one part of the new evaluation model.

**Competing Factors**

The “incoherence” (Firestone & Shipps, 2005) often created by too many initiatives or factors was not a major concern for District A. However, the interaction between the TEM and other district pushes did contribute to the overall interpretation and implementation of the evaluation model. District A’s greatest non-academic concern appeared to be growth of the student population and the resulting need to find physical space for classrooms. This issue received attention at board meetings and in newspaper articles, and was mentioned by the Superintendent through interview. Modular classrooms had to be leased and a feasibility study considering how to house more students was underway. This factor did not appear to have much effect on interpretation of the TEM, however. A review of District A’s financial status found a budget of around $55M and a fund balance of nearly $18M, so while financial needs brought on by TEM (need to hire more staff) were of concern, District A was not overwhelmed by any other non-academic factor to a degree that would take away from the TEM.

Academically, the district was working on initiatives related to aligning curriculum and assessment as well as a goal focused on all students making growth. These goals did not take away from the TEM either, but rather, were themselves affected by TEM, taking second place in regard to attention, time, and human resource. As a result of the TEM, the district had to both slow down its other initiatives as well as hire additional staff so as to continue with the original
factors and accommodate for TEM. The interaction results appeared to lead to the opinion central office staff built toward the TEM, especially related to parts that did not seem to be useful or of any benefit. The frustration created by lack of usefulness and benefit interacted with a frustration that stemmed from being forced to slow down perceived useful and important local initiatives to create an even more negative response to the TEM. This more negative response was evidenced through “buffering” (Honig & Hatch, 2004) activities wherein parts of the model were explicitly kept out of the district by central office staff.

Central office staff listed time as the greatest resource issue with TEM. The Superintendent noted that TEM takes so much time that he “can’t have a conversation with a secondary principal without them bringing it up.” In fact, for first time in all of his time in district, both secondary principals had asked for an additional assistant principal for the upcoming year. These requests were based on a concern that there was not enough time for the TEM, other building administrator duties, and the two major local initiatives. Central office staff had heard building principals report spending a great deal of time in their offices writing up observations instead of conversing with teachers or spending time in the classrooms as they used to do. The Superintendent noted that the pervasive nature of the TEM and its tax on time resources had put a hold or slowed down many other district initiatives. He described this as “frustrating” as well, especially because there was no real perceived need for it. While the Danielson rubric provided some benefit, in the grand scheme of things that benefit did not outweigh the cost to other projects. The Superintendent said, “Things that we had identified as important now all of a sudden had to be pushed down the priority list.”

District A’s curriculum and assessment system was designed to align what was being taught with how it was measured, and to ensure measurement led to teacher reflection and further
refinement of instruction based on student results. While this revolved around the State tested curriculum standards, developing the system was a local decision. It was described by the Superintendent as a “huge investment of resources we’re making as a system.” The investment including contracting with a content expert from the outside multiple times (and paying him) as well as hiring a new staff position (ELA supervisor). Both the Superintendent and Assistant Superintendent noted that without the time requirements of TEM on building principals, the new position probably would not have been needed, as principals and the Assistant Superintendent would have had the time to focus on the curriculum and assessment system.

District A also focused on a goal related to individualizing instruction for all students driven by the idea of growth for all students. As with the curriculum and assessment goal, this was a local decision not directly mandated by the State. The Superintendent noted that data found basic and below basic students were making growth but proficient and advanced students were not progressing nearly as much as they should. The goal was to help address that concern. During the course of the 2014/2015 school year, five to six administrative meetings (with all building and district administration) designed to determine district values and principles related to this goal took place. In addition, administrators were asked to read books and give thought and reflection to the growth concept. Central office staff pointed out that this time had to be split with all the TEM requirements, and the development of the goal moved much more slowly as a result.

Interpretation of the TEM was influenced by concerns with the teaching staff, but not due to pressure from an organized union as much as a central office desire to shield the teaching staff from useless work. No concern whatsoever was expressed by the teacher union at any time; however, the Superintendent asked, “Why would there be?” He found that the TEM “protects
teachers more than the prior model.” The Superintendent worked with the union on some of the gray areas in the model such as how certain data pieces would look (teacher specific data). His overall opinion was that the data side of the model did not contribute anything positive to the evaluation, and at the same time, took a lot of time to put together. As a result, he worked to minimize parts of the TEM. In an example of what Honig and Hatch (2004) describe as buffering, wherein central office staff actively work to keep certain reform concepts from entering the district, the Superintendent and Assistant Superintendent described how one component of the TEM, teacher specific data, was outlined one way by the PDE; only to be wholly revised just weeks before the 2014/2015 school year when it was to be implemented. The change called for additional data-based factors (such as overall achievement and IEP goal growth) to be taken into account, which cumulatively, contributed very little to the overall model in terms of weight. The superintendent said, “think about that, the amount of work that would go into it to make sure the data is right. You’re talking about hours and hours of work for one percent of a teacher’s overall evaluation, which can never be unsatisfactory anyway because of the SPP. So, yea, I’m frustrated. So, we said, we’re not doing that, that is a complete waste of time.” He went on to say, “I think this is ridiculous, there is no value added from my perspective, just something to check off.” In this example, the district buffered one of the components of the model based on the lack of benefit as well as the cost in terms of time. This type of behavior seems to have helped to keep political pressure from the union at a minimal level. As well, because of the strong SPPs in the district, teachers could not have an unsatisfactory rating overall anyway, so there was no real reason for the union to express concern.
Competing factors did not create the “incoherence” among accountability demands that Firestone and Shipps (2005) discussed. However, the time demands that TEM placed on the district had a meaningful effect on other factors that were thought locally to be more important. As these other factors were slowed and the TEM increasingly found to provide very limited gain for its cost, buffering-linked behaviors from the central office emerged.

**Past Experiences**

Honig (2003), Spillane (2000), and Spillane, Reiser, and Gomez (2006) point out that the amount of experience and knowledge associated with a reform influences how likely implementation is to go beyond a surface level attempt. In the case of District A, the experience of the TEM proved to be very different from other reforms. The components of the TEM were also new, including the detail and specificity of the Danielson rubric as well as the complexity of the multiple measures data side of the model. It appeared as if the different experience that came with the TEM as well as the complexity of the multiple measures did play some role in the intent to implement. However, as found in the section below on personal beliefs, the newness and complexity of the observation side of the model did not seem to distract the district from pursuing the original purpose of the Danielson model, suggesting that interpretation was less influenced by past experience than it was personal beliefs.

The Superintendent had been through a number of reforms both predating and following TEM, including standards based reform and, even more recently, changes to the child protection laws. He described the TEM as fundamentally different than other reforms, however. Talking about the structure of public education in the state, with its 500 school districts and vast local differences; he thought that in the past, the State provided enough flexibility in reforms so that districts could fit them to the local context. In addition, he thought that other reforms, based on
their flexibility in implementation as well as their focus, did not seem nearly as all encompassing and rigid as TEM. The Superintendent described the TEM as being “imposed upon us. It felt like we were put upon by the state.” He noted the topic of supervision and evaluation, and how closely the model is tied to instructional behaviors and student outcomes, explaining that it “gets to the core of who we are as professionals.”

Specific knowledge of the components of TEM was limited prior to roll out of the model. While the Superintendent had mentioned that the Danielson model had been in the district for the last 15 years; the Assistant Superintendent did not know the domains or components well prior to the TEM. On the multiple measure side of the model, a school metric (the SPP) as a comprehensive view of data was a new idea, as was elective data (SLOs). Value added modeling (PVAAS) had been applied for some time, but the link between growth and teaching was not totally understood. Training by the local IU, webinars, and other; as well as manuals and the Act 82 regulations all allowed for understanding.

It was not as much the details of the TEM that influenced interpretation as it was the nature of the reform itself. Unlike reforms that had come before, the TEM seemed pervasive and authoritarian. In and of itself, these characteristics were not embraced in District A, but as they mixed with other factors such as a perceived lack of value and personal beliefs, these characteristics contributed to the overall distaste for the model.

**Personal Beliefs**

Central office staff in District A expressed strong individual beliefs as to the purpose of teacher supervision and evaluation related to reflection, growth, and improvement; the need for a system to identify and deal with under-performing staff; and the idea that local control was a better way to manage than was state-driven policy. Collectively, the district had an
organizational set of norms and expectations for operation linked to taking time, being prepared, including stakeholders in decisions, and engaging in organizational behaviors that provided some sort of perceived value added to the district. A portion of the TEM – the Danielson observation rubric – aligned with reflection, growth, and improvement, and as such was embraced by the central office. The other portion of the TEM – the multiple measures data – was in stark contrast to the district’s organizational norms, and as such was minimized (buffered) by the central office.

The superintendent noted how supervision had evolved over the 23 years he had been in education. Early in his teaching career, supervision and evaluation involved just an infrequent formal observation with limited feedback as to strengths and needs. After changing districts midway through his career; however, he began to see the advantages of allotting more time and informal instances of classroom observations. When principals engaged in conversations with teachers about what they had observed, reflection on what was going well and what was not could be had and teachers could focus on improving. The superintendent of District A stated, “There is a lot more feedback now, and feedback is the most important thing. Feedback, in general, no matter what you’re doing in life, is critical to growth. Feedback helps people evaluate where they are moving forward.”

The Assistant Superintendent in District A was even more focused on growth through supervision. Her background as an instructional coach seemed to play a role in her perspective. As a coach, she worked with many different teachers and focused solely on improving instruction (as opposed to evaluation). As such, she engaged in many conversations with classroom teachers looking to identify strengths and needs. Over the course of her coaching, she noted how different each teacher and classroom was, and how supervision and continuous improvement needed to be tailored to each individual. The Assistant Superintendent stated that
any supervision and evaluation system should “encourage teachers to be reflective,” and ultimately, “the point of supervision and evaluation is to help people grow.”

Both the Superintendent and Assistant Superintendent made point of the need for a supervision and evaluation plan to be useful in dealing with under-performing teachers. The Assistant Superintendent phrased this as “allowing the district to put the right people in the right places” or determine that the district was not the right place at all for a person. The Superintendent discussed how there were always a small number of teachers who struggled. While informally identifying these teachers was never difficult, moving forward on a formal plan to remediate or potentially dismiss needed a supervision and evaluation plan that was up to the task of identifying the specific areas of need, documenting growth or lack thereof, and providing a summative rating that would justify a move on the district’s part. As described in detail under the perception of value and worth section (above), the TEM failed in almost every way in regard to this personal belief. The district had no confidence that the TEM would allow underperforming teachers to be addressed, so much so that the district had actually abandoned the model in such cases and used a progressive discipline approach.

The roll out of the TEM from the state level to districts appeared to violate the collective beliefs of District A. The Superintendent and Assistant Superintendent described the district as taking its time before jumping into things, ensuring initiatives are well thought out, communicating and discussing plans, and supporting implementation. Both interviewees referenced the positive and trusting relationship between the administration and the teachers. These were beliefs of the individuals involved, and by extension, of the organization as a whole. The district felt rushed and often had to engage in a mandated process even before clarity from the State was provided in regard to how that process was to look. The Superintendent, referring
to a major August 2014 change to the way one component (Teacher Specific Data) would be applied during the 2014/2015 school year, as well as evaluation forms that were not made public before the school year started, said, “Put yourself in the teacher’s shoes, I don’t know what the reform looks like, I’m not exactly sure what all the components are, I have no control over the SPP score; how would you feel as a teacher? That’s not how we do things around here.” The Assistant Superintendent, referring to the same concerns and her role in making the model work despite these issues, described her own feelings as “frustrated, very, very frustrated; it made me feel like I was put in a position, if our administration didn’t have the positive relationship with the association and teacher, [it would have] caused them to not trust us.” These events and associated personal responses, led to a significant buffering event in District A; wherein the required changes were put to the side for the entire 2014/2015 school year. Despite guidance from the PDE, the district, in consultation with its solicitor, decided the values related to proceeding in a thoughtful and planned way, and maintaining a trusting relationship with the professional staff outweighed what the State had said must take place.

A final personal belief that appeared to emerge was that of local control. Both the Superintendent and Assistant Superintendent expressed frustration over the need to shift away from local goals such as curriculum and assessment systems and growth for all students to a new and intensive evaluation model, especially when the prior evaluation model was doing a fine job. The Superintendent said, “all of a sudden we had this system imposed on us.” The Assistant Superintendent asked, “If a district is high performing and felt our observation and evaluation process was working, why did we need to change?” The Superintendent thought that teachers should be accountable but there would never be a model that could fit the diversity across the 500 school districts in the state. Rather, he expressed a desire for local administrators to be
charged with ensuring teachers were accountable to their students and learning. Detailing what local control might look like, the Superintendent talked about his role and that of his Assistant Superintendent. As Superintendent, he saw his job as knowing what the reform was and how it was working in the district, but also working outside the district to know how it was working for others and what needs to be addressed at a State level. As well, the Superintendent’s job in reform was to figure out what to do as prescribed (Honig and Hatch’s bridging) and what to buffer. He asked things like, “Do we have to do this, what is the value in doing this, why are we doing this?”

Interpreting the SLO process exemplified the belief in local control. The Superintendent described the original presentation of SLOs as something that could be very complex and difficult to manage. In his interpretation, the purpose of SLOs was to “look at how kids were doing – so we didn’t want to put much else there, we just used our curriculum and assessment system that we were already trying to focus on.” By linking the new information to that which already existed, District A “bridged,” (Honig & Hatch, 2004) or assimilated (Spillane, Reiser, & Gomez, 2006) their existing curriculum and assessment system to the new SLO process. Considering the last minute changes to teacher specific data and the “mess” it became, according the Superintendent, the district decided to not follow through with the changes and rather revert to a more simple approach of using the observation score as filler (buffer). Here, in a local control move, the new information was completely buffered, and kept away from the local context.

Personal and organizational beliefs appeared to be a strong factor influencing how components of the TEM were interpreted. In areas where beliefs and the TEM aligned, the component was embraced at the central office level and moved toward implementation in a way
that mirrored both the intent and prescribed process of the TEM. However, in areas where beliefs and the TEM were in contradiction, the district worked hard to buffer the outside influence so as to maintain local values and control.

**Social Interactions**

Social interactions served two purposes to District A. Many of the conversations in which the Superintendent was engaged focused on the larger impact of the TEM, while most of the conversations in which the Assistant Superintendent was engaged focused on how to implement specific components of the TEM. Evidence emerged to support the idea that the worldview of a social group influenced, or at least supported, the interpretation of the Superintendent, and that quality social interactions among the Assistant Superintendent’s peer group were beneficial as it related to local interpretation of specific components of the TEM (Coburn, 2001).

The Superintendent noted that when the TEM was first introduced, superintendents as a group embraced the idea of accountability. The Superintendent of District A made a point of saying, “This can get lost, nothing is more important than the teacher in the classroom, we all agreed on this.” However, once the details of the model, especially the data side with its overreliance on high stakes test results, came to light, concern among the superintendent group started to grow. Overtime, the Superintendent of District A heard colleagues say things like, “I need to protect my principals” and “this makes little sense.” After two years of implementation, talk among superintendents shifted to the utility of the TEM. A recent conversation about dismissal of a teacher in a neighboring district found that most superintendents had begun to use the progressive discipline approach to address poor performance as the TEM was proving to be completely ineffective at removing underperforming staff. Pointing to the irony that
superintendents actually had to go around the TEM in order to dismiss staff, the Superintendent of District A rhetorically asked, “How crazy is that?” Cumulatively, the worldview of the superintendents with which the Superintendent of District A interacted was in contradiction to the amount of time, human resource, and lack of perceived value or worth the TEM brought about.

Other conversations focused on how to interpret and design implementation of the TEM. The Assistant Superintendent identified those who work in curriculum in other districts as well as staff from the local IU as people with whom she had talked. Describing her social circle, she said, “We are blessed in the local area, we have lots of strong people around I can call anytime I need information. Everyone is willing to share and take advantage of each other. Everyone respects everyone’s decisions even when we do things differently.” The Assistant Superintendent thought that discussions among peers allowed her an understanding of what everyone else was doing, which “helps me to clarify my understanding as well as to make decisions in the district for how it might impact my staff.” In one example, the SLO process was being discussed with other districts wondering how exactly to interpret nuances such as setting performance indicator criteria. One district decided to set the criteria at the central office level and allow no teacher input. The justifications for this as well as the consequences were all discussed at a curriculum meeting across many districts. The Assistant Superintendent for District A was able to listen to the interpretation as well as the fallout, bring that information back to District A, and ultimately guide District A in another direction. Such an example demonstrates how greater in-depth processing of information can take place with a group (Coburn, 2001).
As a whole, the social interactions of District A’s central office did not appear to be a major contributor to interpretation of the TEM; however, there were indicators of social influence. These included the collective worldview of superintendents with which the Superintendent of District A interacted as well as the trust and in-depth discussions among those implementing the TEMs in their own districts.

**Interaction of Factors in District A**

The perception of value or worth (Seashore-Louis & Dentler, 1998) and personal beliefs (Firestone & Shipps, 2005) were the most influential factors guiding interpretation of the TEM in District A. The interaction between these factors was marked. The observation component of the model was judged to be important as it aligned with the personal beliefs of teacher reflection, growth, and improvement. The observation component also overlapped with other important district initiatives and allowed certain concepts to be more fully developed as a result. Consequently, District A embraced the observation component, making almost no change to its original design/process and holding true to the original purpose. This commitment was maintained despite the tax the observation model placed on principal time and the resulting slowdown that was felt in other, locally driven, initiatives. This course of events shows how the worth or benefit (importance use, gain) (Seashore-Louis & Dentler, 2005), interacted with personal beliefs (Firestone & Shipps, 2005) and competing factors (Firestone & Shipps, 2005) in a positive way and led to an interpretation closely aligned to the intent of observation component of the reform.

The multiple measures data components of the model were found to provide no use or gain in District A. Moreover, these components violated beliefs related to under-performing teachers and organizational norms associated with understanding initiatives before implementing
them, collectively including stakeholders in decisions, and focusing on a value added. The interaction with factors related to past experience (wherein reforms had not been as pervasive and authoritarian) only compounded local distaste for at least the multiple measure data portion of the model. The combination of these factors led to a great deal of buffering as District A dealt with the data components on the model, wherein the district explicitly worked to limit the impact and met only the most minimal of process requirements. The lack of perceived value or benefit and violation of beliefs seemed to exacerbate the concern expressed in regard to other factors in the district. While the time it took to engage in the observation model, and the resulting slowdown of other initiatives, was more or less accepted in regard to observation, the time involved in the data side of the model was identified as a major source of frustration due to the slow down it caused with those same initiatives.

**District B**

**Background.** District B was in a rural fringe locale (National Center for Educational Statistics, 2015). The district was 84 square miles in size, with 3140 students attending six schools, with school performance profile (SPP) scores that ranged from 72.8 to 88.2. Twenty percent of the student population was classified as economically disadvantaged. The annual budget for District B was about $48 million and data from the state Department of Education disclosed there was around an $8 million fund balance (PDE, 2015). The district website listed 16 central office administrators and 10 building administrators.

A review of documents suggested there were some meaningful issues that could influence administration in the district. School board meeting minutes found non-unanimous votes on topics related to sports coaches, driver’s education and biology curriculum, high school activities and clubs, and the superintendent’s salary for the 2014/2015 school year. One of the biggest
issues was a petition from parents in a neighboring township to leave their district and join District B. A School Board resolution adopted in January of 2015 provided a timeline of events and formally noted the district’s position on the issue. The petition was filed in July of 2012. From 2012 through June of 2015 local courts and the Department of Education addressed issues with the request. District B conducted a number of studies to determine the impact the change would have. While the income from the additional township would offset the costs in staff and supplies, District B would need to take on a number of building renovations to account for the additional 300 students that would join. The Board resolution officially took a neutral stance on the issue - not encouraging or discouraging the Secretary of Education from allowing the township to switch districts.

The Superintendent and Assistant Superintendent for academics were interviewed for the study. The Superintendent had been in his position for three years. Prior, he had worked for eight years as a superintendent in another district and for three years as an assistant superintendent. He had worked as an elementary principal and teacher as well. The Superintendent had 27 years in the education field with 21 of those in administration. The assistant superintendent had been in his position for six years and was a middle school assistant principal prior – all in District B. Before that, he worked as an assistant principal and classroom teacher.

**Perceptions of Value or Worth.** While the perception of value or worth factors of importance, usefulness, and gain (Seashore-Louis & Dentler, 1998) were mixed in District B, the overwhelming weight of evidence suggested the benefit associated with the teacher effectiveness model (TEM) led to an interpretation guided by genuine desire to match the underlying purpose of the policy behind the model (Honig, 2003; Spillane, 2000). Results found that the supervision
and evaluation model prior to the TEM was not all that useful and did not provide much benefit for the district. Alternatively, the TEM provided an improved teacher supervision and evaluation, a topic that was important yet underserved. Additionally, the model was useful as a point of leverage for other initiatives that the district was able to bridge directly to TEM. Some concerns with the model were noted, but these did not outweigh the advantages.

Central office staff found no use or benefit in the prior evaluation model. The Superintendent described the older model as a “snapshot.” There was a pre-conference, a formal and planned classroom observation, and a post-conference. The evaluator would wrap up the process with an evaluation report, which would go on file. The Superintendent commented that this, “didn’t really lead to anything different in terms of teaching.” He empathized with teachers, noting, “You have to hope it’s a good day and the kids are engaged in the lesson – so not on a sunny day in the spring or right before a holiday.” The Assistant Superintendent described such infrequent observation as “isolated” and limited to a single point in time. He noted that “there was no reflection; it was done to you as a teacher, a passive process. You just make sure you had a lesson plan and it was in your binder when the administrator came in.” A noted side effect to such was a lack of a differentiated supervision model. Because the formal model was so easy and effortless for teachers, there was very little motivation to engage in an alternative model that required more work. As such, central office staff saw little personal growth among teachers.

Another problem with the old model was its subjectivity. Conclusions were often very general, such as “She doesn’t plan very well,” but rarely was there supporting evidence to back up a claim. As a result, a poor evaluation often became an issue with the teacher’s association. The Assistant Superintendent noted that improvement plans that stemmed from union meetings were often very technical. The evaluation of plan completion usually had to do with whether or
not something was done (read a book, attend a professional development training), not whether or not the teacher’s instruction actually improved. He noted, “They would do the task to satisfy the improvement plan but we wouldn’t necessarily see an improvement in practice.” The Superintendent gave an example of an under-performing teacher whose principal requested the district not grant tenure at the at the end of the year. Unfortunately, there was no evidence to support the decision, just the principal’s opinion. As a result, the teacher had to be granted tenure and as the Superintendent noted, “We’ll have to deal with 32 years of ineffectiveness.”

The TEM provided a more useful evaluation through reflection, objectivity and accountability. Central office staff reported more conversations between principals and teachers about instruction and what planning looks like. The Superintendent noted, “In the past, no one thought about planning, but now we have a rubric that looks at it.” While he conceded the observation model was a cumbersome and time consuming process, he “loved” the conversations that were taking place in the schools and classrooms. He thought, “These conversations are truly helping teachers become better teachers.” The Assistant Superintendent agreed, noting, “I see significant change in how people are planning, it’s much better than before, thought-out and reasoned.”

Use of the Danielson framework brought objectivity, consistency, and accountability to the supervision and evaluation process. The framework allowed evaluators to be much more specific in their feedback, citing specific Danielson domains and components. Moreover, teachers were forced to provide evidence across the four domains, creating an active and ongoing system. The Superintendent reported, “There is a greater onus to gather information, a greater onus for the principal to know his or her teachers.” When administrators expressed concern with teachers, they were able to do so in a very detailed way and cite specific evidence that did or did
not exist. As a result, the debates with the association that would stem from the older subjective model seemed to disappear. Discussing meetings with under-performing teachers during TEM implementation, the Assistant Superintendent described the conversations as, “here it is, the evidence is right here. In the past we would make a summary statement but we couldn’t really prove it, now the evidence is there.” The Superintendent talked about the inclusion of student data leading to intensified levels of accountability, “Let’s face it, teachers are being evaluated not only on their performance but on student performance, our principals are being evaluated not only on their performance, but on teacher performance and student performance. So, all of a sudden there is a heightened level of accountability and concern.”

The TEM allowed District B to leverage many of the curricular initiatives that had struggled prior to its implementation. The district had not been able to establish a differentiated supervision model before the TEM, mostly because the work involved in such a model far outweighed the older passive observation system. However, with the TEM and its focus on gathering evidence and artifacts to prove proficiency across the four domains and their respective components, signing up for differentiated supervision actually presented an option that was less time intensive. This was a major advantage that stemmed from the TEM, especially as it interacted with other initiatives in the district and the personal beliefs of staff in the central office. The Assistant Superintendent talked at length about outcomes from the new differentiated supervision. “We saw teachers doing book studies that were reflective, data teams truly looking at students, high school teachers looking at literacy, and considerations of what interventions were working.” Prior to the TEM the district had struggled to understand assessment and how assessment was associated with teaching and learning. The student learning objectives (SLO) portion of the TEM served to change this. The Assistant Superintendent noted,
“The TEM forced some teachers who had no interest in student data before to pause, and say maybe I should pay attention to this, and ask very different questions.” The SLOs “forced us to focus on student learning.” He talked about conversations never had before regarding growth rates, good assessments versus bad, and noted, “What’s going on, without the TEM I’m not sure it would have happened on its own. It’s focused our attention.” Moreover, the Assistant Superintendent said, “We’ve tried to do this before but couldn’t get staff to move in that direction. TEM has allowed us to go to these things.”

The TEM was not without concern in District B. Relative to teaching staff, the Superintendent noted an incredible increase in stress and anxiety due to both the grand scale of work involved in the collection of evidence as well as the major shift in attention to student growth and outcomes. He reported, “I’ve never seen teachers more stressed out than I have the last two years. And anxiety translates down to students, that concerns me.” Relative to administration, the concern was with a displacement of goals (see Weber, 1922). The Assistant Superintendent expressed concern that because the model was so evidence based, the emphasis had amplified the process of gathering evidence instead of looking at the whole picture of instruction. He noted, “The feeling of a class, the pace of a class, the interactions, it all gets lost because principals are too busy just gathering evidence.”

The district’s capacity made parts of the TEM difficult. One concern for administrators was management of all the data. This was described as “almost overwhelming, even with a software system, efficacy is lost in the volume.” To make this point, he noted that teachers could provide between one and three pieces of evidence across all the components in domains one and four of the model, leading to as many as thirty or more sources of evidence for a principal to review – per teacher. “Just for the principals to review the evidence takes hours and hours and
hours. There is no way a principal is going to look at all of this stuff.” A final concern with the TEM, also related to capacity, had to do with the inability of the district to take advantage of all of the opportunities the TEM seemed to be creating. In example, as differentiated supervision took hold, there was a need to tailor professional development for many small groups. However, the district did not have the staff or the time to plan each of these events nor the ability to find “experts” in so many different topics.

Despite the improvements with TEM over the prior model, the TEM still did not offer a viable way to work toward dismissing under-performing teachers. The Superintendent suggested that the inclusion of school performance data was a major inhibitor. In example he offered, “I have an ineffective teacher in a building with an 88 SPP. That teacher could score the lowest possible in all those domains but will never be rated unsatisfactory.” To remedy this, he noted that the teacher would need an improvement plan and administrators “better be darn sure to turn up the heat, otherwise she’ll sit there for years and years [without being dismissed] just because of the building she’s assigned to, not because of her instruction.”

Evidence supported the conclusion that the change brought about through the TEM provided the importance, use, and gain described by Seashore-Louis & Dentler (1998). Holding teacher’s accountable, facilitating teacher growth, and allowing the district to make evidence-supported objective decisions regarding evaluation were all important to central office staff. These things were not possible with the prior evaluation model and the TEM allowed District B to move forward with each. Moreover, the TEM was used as leverage by the central office to push other district initiatives. While some concerns were noted, and solutions to these appeared exceptionally difficult to find, central office staff judged the TEM to have great value in District B.
Competing Factors. Firestone and Shipps (2005) wrote of competing factors from the perspective wherein central office staff would be overwhelmed and under-resourced by additional outside mandates. As a result, mandates would receive limited attention and resources, and the initiative would struggle for an implementation no greater than minimal levels of compliance. In District B the competing factors had just the opposite effect. Central office staff bridged (Honig & Hatch, 2004) components of the TEM to existing initiatives related to instruction, and separated central office staff roles to reduce the interaction of non-instructional initiatives on TEM interpretation and implementation.

Improving instruction was an important factor in District B. The Superintendent described the district as very “1985-like,” meaning that the district still operated as if it was 1985 in terms of buildings, curriculum, assessment, and technology. He stated, “There was nothing wrong with 1985, it was great for 1985, but not so great for 2015, when we still use the same technology, tests, etc. as we did in 1985.” To move forward, the district, through the Assistant Superintendent, focused on 21st century skills including Science, Technology, Engineering, and Math (STEM) curriculum as well as literacy across the curriculum at the high school level. The district was working hard to “re-build” after recent budget cuts had slashed staff and programming. “We have kids leaving for college going pre-med having never had an anatomy course because that was something that was cut.” The Superintendent had managed to build five new full time teaching positions into the 2015/2016 budget, and it was clear that his goal was not just to survive, but to move District B towards something much greater, “We’re trying to rebuild and get back to offering programming that does more than just what Chapter Four requires.”

Instead of seeing the move toward 21st century instruction and the TEM as competitors for limited resource, District B carefully wove the two together, using the TEM and its legal
mandate to leverage improved curriculum, instruction, and assessment. The bridging between the TEM and the instructional initiative, and resulting interaction between the TEM and district values as well as other factors demonstrated how under certain conditions, a new reform was embraced and endorsed with great enthusiasm, even in a local environment already charged with its own initiatives.

The consistency of process across the district was another factor in District B. Prior to the current superintendent taking his position, all the schools “worked as their own fiefdoms” with no standardization between them. As a result, everything from the hiring process to teacher evaluation to budgeting was somewhat different and often outside of acceptable standards. The District was trying to develop common and defendable procedures, “We are taking steps to improve processes, to be a financially, legally, and academically sound organization.” Another factor listed was facilities. The Superintendent noted to the interviewer, “The building that you’re sitting in was built in 1914 – and little has been done to it since, it’s falling apart.” Over the past few years, the district had renovated two elementary schools and was in the process of renovating a vacant building in order to move the central office. The Superintendent noted that he took responsibility for these latter two factors, improving processes and updated facilities. He charged the Assistant Superintendent with all things academic, “He is my chief academic officer, my go-to person. He answers all the questions.”

District B did well to manage competing factors and the TEM. In the case of instructional initiatives, the District bridged the new information from the TEM and used the State mandate as a way to leverage already existing pushes for instructional change. Indeed, it appeared that the perception of value or worth associated with the TEM was enhanced by the influence of the TEM’s overlap with instruction related emphases. In other cases, such as
attempts to improve processes and facilities, the District separated duties among central office staff in order to avoid limiting time and attention resources. Other factors, such as a fiscally conservative community, were addressed through a strategic plan, which was used to justify expenses. While teacher concern was noted, there was no evidence to suggest that this factor changed the way the TEM was interpreted in any way.

**Past Experiences.** The experience of TEM as a mandated reform seemed similar to most others in District B. As such, the confusion and frustration that accompanied the TEM did not seem to create a meaningful barrier to interpretation. Central office staff in District B were familiar with the components of the TEM prior to its roll out. This familiarity, when paired with the usefulness of the TEM to leverage other district initiatives, strengthened the embrace of the model in District B.

While the history of reform was not thought of in a positive light, central office staff seemed to accept reforms as par for the course. Staff would focus more on how they could use the reform as leverage more than how they could keep reform ideas out of the district. The Superintendent described the process of reform from the state as coming in a “ready, fire, aim” manner, and noted that the TEM was no different. He reported that the State never seemed to think of the unintended consequences nor did they have a plan to react to problems once they appeared. He cited the roll out of recent Child Protection regulations and background checks and clearances as an example of a different reform presented in the same way, brought about with good intentions but in practice woefully unsupported or thought-out. Despite the frustrating approach, the Superintendent seemed to take it all in stride, and he found the benefits the TEM brought to District B far outweighed the drawbacks, including those which presented as a result of poor planning on the State level.
The Assistant Superintendent had a solid understanding of the components of the TEM. The Danielson rubric had been used to an extent in the district for some time, and while additional training and study was needed as part of the TEM, the general ideas and domains were not new information. Value added data through PVAAS had been accessible and reviewed for many years prior to the TEM; and while the link between teachers and student growth was new, the concept of student growth and how PVAAS calculated this was familiar. The SLO process was new only in “vocabulary and process.” The idea of using local tests and measures to gauge student performance and growth was a hallmark of the Assistant Superintendent’s belief system and had been a major focus on the districts attempts to “move out of 1985.” The components of TEM, both familiar and aligned with personal beliefs, probably increased the likelihood of each being embraced in this district.

**Personal Beliefs.** Three major personal beliefs emerged from the evidence. Two of these beliefs, evaluation as an ongoing process and sincere teaching and learning, were used to bridge the TEM into the district. It appeared as if these personal beliefs drove the initiatives District B had been attempting to implement for years, and the legally mandated nature of the TEM created just the leverage central office staff needed. The other belief that emerged from the study, teacher collaboration, was supported by the TEM as it allowed more teachers to pursue differentiated supervision; however, TEM components linked to student outcome data also limited collaboration by creating competition within the district.

The first belief focused on the idea of evaluation as a comprehensive process, not simply a “snapshot” in time. The Superintendent of District B said, “Evaluation should not be a snapshot; it should be a comprehensive picture of the teacher in a multitude of interactions with students” and noted, “I don’t sit out and watch my secretary for 45 minutes one time a year, we
laugh at that, but we don’t laugh when we think that’s what we’ve been doing with teachers.”

He noted that single and isolated observations allowed “really bad teachers to be really good actors,” putting on a show for their observation but spending most of the school year engaged in less than desirable teaching behaviors. The Superintendent was able to recall communication from parents expressing concerns about how a teacher interacted with students and parents, yet he felt he had no choice but to give a satisfactory rating because the teacher had done well during the single event when she had been observed. The Assistant Superintendent of District B called this a “dog and pony show.” He said, “We did all announced visits so it was always just a dog and pony show. It really did nothing to promote ongoing learning or self-reflection, and there was very little analyzing student assessment data.” Central office staff expressed a desire to see teachers in an ongoing manner, with principals spending time in classrooms throughout the school year and bringing in data and evidence that was comprehensive to the teaching process. While an increase in classroom time for building administrators had “evolved” over time, and not just appeared with the TEM, the ongoing nature and all-encompassing evidence gathering components of the TEM aligned well with this personal belief in District B.

Another major personal belief had to do with teacher collaboration. Unlike the belief in ongoing evaluation, which was aligned with the TEM, the weight given to student data and especially high stakes test results in the TEM was cited as contrary to this belief. At the conclusion of the 2014/2015 school year, when teachers first began to understand the impact of student scores on their evaluations, there was emerging concern with competition between teachers, grade levels, and schools; as well as a shifting culture of criticism toward anyone thought to contribute to poor outcomes. The Assistant Superintendent told about teachers in the high school telling content teachers (English, Algebra, and Biology – the tested subjects) to “do a
better job next year” because the scores were bringing everyone’s evaluation down (through the SPP). He also reflected how some teachers no longer wanted to “share” their students, keeping them in their own classrooms for all instruction as opposed to sending them out to reading or special education specialists. This bred a culture of mistrust and led the Assistant Superintendent to say, “the collaborative piece is limited by high stakes test data.”

The final personal belief that emerged through analysis of the data collected was a focus on sincere teacher and student learning. During a report to the School Board during the 2014/2015 school year the Assistant Superintendent said, “We’ve had a laissez-faire attitude about assessments tests, we’re not incorporating things into the curriculum to align with testing subjects” (Kline, 2014). Both the Superintendent and Assistant Superintendent referred to making improvements in the curriculum, instruction, and assessment processes in the district, and shied away from endorsing “games” by which scores could be manipulated, preferring rather to work toward sustainable improvement through improved instruction. This belief was also evidenced through the way staff talked about what was important in the district. For instance, the Superintendent listed curriculum, standardization of processes, and facilities as the three most important issues going on in the district. He did not mention a major movement by a neighboring township to succeed from their school district and join District B. This move had been filed with the State and the PDE was on the verge of making a decision, yet the Superintendent chose to focus on topics directly related to improving teaching and learning. In District B, this belief in genuine learning was leveraged by the TEM. Many of the initiatives borne of the central office over the past three years had gained little traction; however, as outlined in the section on the perception of value and worth above, District B was able to bridge
TEM components like a differentiated supervision plan and the SLOs to these district initiatives, all of which were aligned with the focus on genuine teacher and student learning.

**Social Interactions.** The social interactions surrounding the central office addressed two themes, specifically the amount of work that came with the TEM and the frustration over a lack of clarity and support from the state level. Interestingly, while those with whom the central office staff were interacting noted a number of problems in this regard, District B did not appear to change its interpretation of the TEM as a result. Commenting on conversations with his staff, the Superintendent reflected hearing things like, “I have no time to do this. What is going to be sacrificed?” Talking with other superintendents led to similar; “While everyone is implementing TEM differently, we all talk about the same struggles. We hate the process, mostly how much time it consumes.” Interaction related to this topic seemed to reinforce what District B staff were sensing locally; however, it did not appear to have much of an impact on the interpretation of TEM overall. As the superintendent stated, “It’s like any job, you bitch about what you have to do and then go do it anyway.”

There was a great deal of ambiguity with parts of the TEM, and the questions that arose as a result were often discussed within the social networks of District B central office staff. The Superintendent referenced the hasty roll out of the model and an apparent lack of attention to detail, noting, “When you call the State there is no one there, or there is no one willing to help, they just told us it was a local decision.” The Assistant Superintendent discussed a number of specific issues where there was no clear guidance. Working with colleagues in other district, who all faced the same questions of clarity, the social network debated topics such as how many students a teacher should have on a case load in order to have student achievement and growth data attributed to them. This concept was part of the TEM through PVAAS as well as SLO;
however, the State would not provide a definitive N count to districts, only “referring administrators to their legal counsel.” Assistant superintendents talked about what N count seemed most appropriate, and under what circumstances a particular N count might be used. Often, this discussion led to more confusion, as the Assistant Superintendent from District B said, “As similar as we are, there are a lot of different interpretations. We would get to a point where we thought we knew what to do, then hear from someone doing it differently and we would rethink the whole thing.” Another topic often discussed was how to designate certain staff. Professionals such as reading specialists, librarians, transition coordinators and others did not fit into any of the PDE provided evaluation rubrics. “We had no guidance from the State so we had to just figure it out amongst ourselves,” stated the Assistant Superintendent.

Often, through monthly meetings at the IU, list-serve emails, or simply calling one another, the network of central office administrators charged with interpreting the TEM would discuss what each thought and weigh other’s thoughts in order to make a local decisions. One such example demonstrates the link between the influence of the social network and the personal beliefs found in District B. In order to complete an evaluation, an SPP score must be attributed to a teacher. For most teachers, the evaluation is completed during the fall of the following school year, so observation and student outcome data from a single year can all be compiled into the same evaluation. However, for some teachers, such as those who are in line to be awarded tenure following the completion of their third year, waiting until the fall is not an option. The social network referenced by District B staff discussed different ways to deal with this issue, including assigning a SPP score from a different school year simply to have a number to fill in the model. Others in the network were uncomfortable with this idea, though, and instead withheld the SPP data all together, replacing the score with the observation score (administrators
would input the observation score into the SPP “cell” in the calculating spreadsheet, appeasing the models need for a data input yet in effect only using the observation results). District B, at least in part influenced by other districts as well as the personal belief that input in the model should sincerely reflect a teachers actual work, went with this latter approach.

The social networks of central office staff in District B allowed for discussion, venting, and fine-tuning of details not prescribed from the state. The opportunity to work with others was clearly helpful; however, it did not appear to influence interpretation of TEM to a meaningful degree. While many of the implementation details, such as N size or SPP scores, were determined based on conversations, the more general interpretation that embraced the TEM was driven by other factors.

**Interaction of Factors in District B.** Seashore-Louis and Dentler (1988) noted the influence organizational gain played in how reform information was interpreted. In District B, the TEM was perceived by central office staff to allow for a great deal of gain, and as a result, was embraced. This process was not straight forward; however, and did not always lead to the second order change described by Spillane (2000). The observation (Danielson) portion of the TEM offered two potential gains. The first related to teacher growth, especially in regard to planning. The gain for District B encouraged the organization to implement the observation in a way aligned with the original intent (second order change). The other gain that stemmed from the observation rubric was related to teacher response. Because the Danielson was so heavily dependent on gathering evidence, teachers saw the differentiated supervision option as more advantageous. The district had been trying to establish a differentiated supervision model for some time, and while an unintended consequence of the TEM, this reaction provided a gain none-the-less. Another gain was found through the SLO portion of the TEM. The district had
been working on linking assessment to instruction for some time, but had struggled to get the idea to take hold. By bridging (Honig & Hatch, 2004) the SLO to this assessment push, District B was able to gain a great deal of leverage related to their original initiative. While the implementation of the SLO remained more or less aligned with the original purpose of the TEM, the real motivation in the district was linked to this organizational gain. A final noted gain had to do with an overall and general improvement in teacher evaluation. The TEM provided a standardized, objective, and defendable model, something that had not existed in District B prior.

Interaction between other factors in the district multiplied the gains that were realized. Central office staff held strong personal beliefs in many of the areas where they thought the TEM could lead to gain. Beliefs in an ongoing evaluation model, teacher collaboration, and genuine teaching and learning (as opposed to just the appearance of such) linked nicely to perceived outcomes of the TEM including teacher growth, differentiated supervision, using assessment data to drive instruction, and having a more defendable evaluation plan. As well, central office staff had a strong understanding of the concepts that made up the basis of the TEM, including the observation model, PVAAS, and the SLOs. The strong personal beliefs encouraged a desire to bridge the TEM to already existing initiatives and the background knowledge allowed the bridging to be done in an informed and successful manner.

A number of other factors were isolated as much as could be so as to allow the TEM to be implemented. The Superintendent took on factors related to establishing systemic processes across the distinct, rebuilding staff and programming from budget cuts years before, and working with a conservative school board to secure the support needed to fund many changes. Taking these things on allowed the Assistant Superintendent more time to focus on TEM and do the necessary bridging between new TEM-related concepts and pre-existing district initiatives.
Factors related to teacher response and the concerns noted by those in the Superintendent and Assistant Superintendent’s social networks were of note but did not appear to play a meaningful role in how the TEM was interpreted.

**District C**

**Background.** District C was in a small city locale (National Center for Educational Statistics, 2015). The district was 76 square miles in size, with 4956 students attending 10 schools, and school performance profile (SPP) scores that ranged from 75.8 to 94.4. Thirty four percent of the student population was classified as economically disadvantaged. The annual budget for District C was about $70 million and data from the state Department of Education suggested there was a fund balance of about $18.3 million (PDE, 2015). A review of school board minutes found some community concern with overcrowded classrooms and a potential decision in the spring of 2015 to outsource para-educator support. Technology was given a fair amount of time relative to discussion at board meetings, and it also was predominate on the district webpage.

The Superintendent and Director of Curriculum and Instruction (C&I) were interviewed for the study. The Superintendent was entering his 6th year in the position and, prior, had been the Assistant Superintendent in District C for nine years. He also had experience as a middle school principal and was originally a teacher at the middle school level. The Director of C&I was entering her 12th year in that role in District C. Prior, she had worked as an Assistant to the Superintendent in another district for five years and had been an assistant principal for 9 years before that. Originally, she had taught French and Spanish.

**Personal Beliefs.** Firestone and Shipps (2005) reference the potential influence of personal beliefs in how a reform is interpreted, suggesting that the alignment, or lack thereof, of
central office staff beliefs and the reform in question goes a long way in predicting the likelihood of the reform being implemented in a way consistent with the original intent of the policy rather than just its form (Honig, 2003, Spillane, 2000). This played out in District C – where the personal beliefs of the Director of C&I and to a lesser extent that of some members of the school board seemed to be the driving force behind teacher effectiveness. Because of these beliefs, the topic of teacher evaluation, including teacher accountability for student achievement, was important in the local context. Individually, the Director of C&I had a career-long passion for supervision and evaluation, and her experiences had taught her to implement and follow processes as designed. The school board, reflective of the diverse local community, believed in student achievement and that teachers should be evaluated based on how well they effect such achievement.

The Director of C&I believed that the purpose of supervision was to provide teachers with information and feedback that would allow each to improve. The purpose of evaluation was to determine if a teacher’s performance was at a level good enough to justify continued employment, asking “Have you done enough good work to remain?” The Director described her first evaluation as a young teacher, wherein the administrator commented on how well her window blinds were hung and how organized her bookshelf was. She also recalled her first evaluation report, which labeled her class as French when it was actually Spanish. This event exemplified a widespread discrepancy between her personal beliefs and the practice found in her experience. As such, she embraced a career-long focus on supervision and evaluation, stating, “I have a passion for supervision and evaluation not being done right throughout my whole career.”

The Director described a number of attempts to change or update models with mixed results at the implementation level. She saw the Teacher Effectiveness Model (TEM) as a framework
aligned with her personal beliefs and an opportunity to improve the supervision and evaluation process in schools.

The Director of C&I also expressed a personal belief in doing things as prescribed. This belief was driven in part by a desire to be consistent and fair across all staff. She cited an “inherit unfairness” when processes were not tightly controlled and described the result of lax regulation as “bastardizing the system and leading to crap processes.” Describing the job of attributing students to teachers for the PVAAS growth metric, the Director retold how the central office worked out formulas so as to mathematically assign the correct percentages instead of simply allowing individual principals or teachers to derive a percentage based on their own random criteria. Similarly, the central office set the performance indicators for the SLOs across the whole district. This personal belief in tight coupling (see Gamoran & Dreeben, 1986) represented what Spillane, Reiser, & Gomez (2006) referred to as a schema – in that the Director had a mental model for how new initiatives should be implemented. The prescribed approach of the TEM fit this schema perfectly and the alignment, paired with her passion for the topic, created a situation where in the TEM had a great deal of personal value.

The Director of C&I contrasted her tight control schema with how her colleagues in other districts gave teachers much more flexibility. The latter, the Director feared, would lead to inconsistencies and great difficulty justifying “failing” performance if necessary. Such fear, or worry, was another driving motivation behind the Director’s belief in consistent processes. Discussing why she spent so much time making sure the data entry portions of the TEM were correct, she said, “I’m afraid if I don’t go in and look at everyone, the State is going to say I didn’t check that roster. I’m fearful it’s too important to a teacher that it’ll mess with their
evaluation. I probably found 12 to 15 mistakes. Would they have resulted in a failing PVAAS score, I don’t know; but if I don’t correct it, there’s a mistake in there and I know it.”

The influence of the School Board was mentioned by the Superintendent as well as the Director of C&I. The Superintendent described the Board as “strong and professional,” noting that while they trusted the administrative staff to get the job done, they also had an expectation that the job would be done. In 2010, prior to the TEM, a committee from District C developed and took to the Board a plan for supervision and evaluation. Upon review, the Board asked the central office to add to the plan a section detailing how teachers would be held accountable for student achievement. According to the Director of C&I, “At the time we had a local requirement that students had to pass the PSSA to graduate. The Board thought if we are holding kids responsible then we need to be holding teachers responsible, too.”

It is worth noting the Superintendent was not as personally engaged with supervision and evaluation as was the Director of C&I; however, there was nothing in the TEM that violated any of his personal beliefs, either. As a result, he supported TEM implementation. The indifference is important in that, similar to the lack of impact it had on other initiatives in the district (detailed below), the TEM came to represent something that did not interfere with the operation of the district nor offend the Superintendent’s personal beliefs.

The association between the TEM and the personal beliefs of the decision makers in District C represents the most important factor in how the TEM was interpreted. The Director of C&I, who had been given the authority by the Superintendent to take on the TEM initiative, had a passion for supervision and evaluation and a mental understanding (schema) for how initiatives should be structured. The TEM aligned with both of these beliefs, and as a result, was initially
bridged (Honig & Hatch, 2004) into District C with every intention of second order (Spillane, 2000) change.

**Perceptions of Value or Worth.** The perception of value or worth of the TEM related to how important the topics of supervision and evaluation were to the Director of C&I, the usefulness of the Danielson model in growing teachers, and the benefit it brought to the district by showing building level administrators what good instruction looked like. Concerns with the TEM revolved around an emerging discrepancy between the opportunities it created and the actual implementation at the building level as well as the inordinate amount of time that went into the multiple measures side of the model.

As is detailed in the section on personal beliefs, supervision and evaluation were very important to the Director of C&I. Prior to the TEM, the Director of C&I had worked hard to move administrators toward a model of supervision that could provide meaningful feedback to teachers. She provided trainings and encouraged the scripting of lessons, the use of detailing instructional observations as opposed to just providing summary statements, and the provision of a summative report. Unfortunately, she had experienced a struggle bridging the gap between her ideals and actual practice. The recommendations and suggestions that would follow an observation were often, “crap, they were just horrible.” Recalling a summative report where a building principal suggested that the teacher should have had students pass out worksheets, as opposed to the teacher doing it; the Director of C&I noted, “This is silly, and looks silly to a teacher. What does that do for instruction? As a teacher, this just says you have a supervisor that isn’t going to help me much. Do I have respect for this person? If this is the silly thing I’m told? There is no supervisory process occurring. The administrator doesn’t know anything about the content or instruction. How silly is that? Absolutely silly.” In District, C and to the
Director of C&I, the TEM presented as a potential framework that might allow for advancement where in the past such could not be realized. This opportunity was perceived to be so clear, and the topic so important, that District C signed up to pilot the observation model two years prior to mandated implementation. The Director of C&I noted, “I was happy with Teacher Effectiveness because it was going in the right direction. I was a proponent because it was similar to what I was trying to do.”

The Danielson model provided some useful results in District C. The Superintendent noted how the TEM provided much more attention to the good teachers in the district and the instruction they were planning and delivering. He noted, “Good teachers, they want us to know they put time and effort into teaching.” Conversely, when working with teachers in need, he reported, “We see problems that were always there, but the TEM helps to show them.” The Director of C&I confirmed the Superintendent’s perspective, noting how the TEM expanded the realm of supervision to Danielson’s four domains. As a result, administrators were able to see “all the work that goes into a lesson plan, all the decisions a teacher is making.” Teachers themselves were observed to be much more reflective, and the post-conferences were much richer, with, “conversations around evidence instead of focusing on process and documents.” As well, teachers were noted to embrace the TEM because it allowed them to rate themselves on the four domains, discuss these ratings with their administrators, and receive targeted and specific feedback. The Director of C&I observed that, “Teachers are learning and growing and becoming better teachers.”

The greatest benefit in District C appeared to be the knowledge the TEM brought to building level administrators relative to what solid instruction looked like. Prior to the TEM and noted above, many administrators were providing little more than “silly results” through the
existing supervision and evaluation model. The TEM, structured and detailed via the Danielson Framework for Teaching, provided building administrators with both a broad four domain framework but also specifics within each so as to understand and know what it was they should be looking for in the classroom. The Director of C&I had found in the past, “If there was something superficial but looked good, that was good enough; but now you can look closer to see how much thought was put into the lesson.” Historically, if a lesson plan was poor, administration would assume the teacher was doing a poor job; but through the TEM, administrators were able to see why a teacher did one thing or another. Additionally, if a teacher was strong in one area, like building relationships with students and their families, but weak in another, like instruction, the TEM allowed administrators to note both – strengths in one area and weaknesses in another.

Two major and one minor concern with the TEM emerged from the data. The growing sense that the opportunity inherit to the model might not be realized in District C and the overwhelming amount of time it took to implement and maintain the full model were meaningful. The limited improvement the TEM provided in regard to addressing underperforming teachers was noted; however, this was not a major concern.

The Superintendent expressed concern as to whether the district would be able to take advantage of all the potential opportunities. The Director of C&I echoed this worry, commenting that the “Danielson model is a lot to learn, the differences between artifacts and evidence, what evidence is good evidence, and how to understand the rubric differences.” She noted that administrators were giving a lot of “distinguished” ratings, and these appeared to be related more to how well liked a teacher was than to the quality of evidence. She noted, “There is a lack of evidence to support the ratings. The principals still don’t understand the model.”
Giving a specific example, she detailed how a staff member had simply posted links to lesson plans in her observation but had earned a distinguished evaluation. “I was very disappointed that the principal didn’t understand the difference, that a document was an artifact, not evidence.” Moreover, the Director of C&I stated that the lesson plans were rather standard – proficient at best but far from earning a distinguished mark. While the Director of C&I had a number of administrative trainings planned to teach more about the Danielson model, she worried that it would end up as many in the past, with “people going through the motions but really just doing what they have already been doing.”

The time involved with bringing the model into the district and setting it up for implementation was described as “overwhelming.” The Director of C&I described it as the “magnitude of work” and despite her passion for supervision and evaluation, she was finding herself turned off by all the work and worrying that it was not justified by the end result. Much of this work was linked to the multiple measures side of the evaluation model. There were a multitude of factors that went into the SPP score for each school in the district, and all factors for every school had to be reviewed before submission to the State. Each student taking a PSSA or Keystone exam had to be attributed to content-related teachers in order for the value added model to correctly compute teacher effect. As described above, the Director of C&I went through each teacher’s roster to ensure correctness. The SLO component also had a “huge process” with a great deal of decision making made at the central office level. The Director explained, “I don’t disagree with the concept of accountability at all, it’s just the magnitude of the process that’s the problem. And what we put into it is not equal to what we get out, so much technical work for so little contribution to the final outcome of the model.” Talking specifically about the value added model, the hours of roster verification, and the resulting feedback that is
reduced to a color-coded score she asked, “Right now we just get a number that’s colored, if it’s one color that’s bad and another color that’s good. Is this giving me any more information on the teacher than I already know from what they’re doing here every day? I’m not sure it makes sense.”

The TEM provided minimal support when dealing with under-performing teachers but also presented some road blocks when it came to dismissal. The Director of C&I was optimistic that the detailed domains and components of the Danielson model would allow for more specific evaluation findings and would link directly to targeted improvement plans. She worried that not all administrators were proficient enough in the Danielson model to make this opportunity a reality, but also noted that prior models of evaluation did not lead to solid improvement plans either. Both the Director of C&I and the Superintendent noted the limitations set on dismissal by adding the data portion of the model into consideration. Since the SPP scores in District C were rather strong, most teachers would be protected from an unsatisfactory rating, even if their instruction was subpar. The Director of C&I told of a teacher who was not doing a very good job; however, her SPP and SLO scores kept her at a proficient level. While somewhat frustrated by the limitations of the TEM, the Director was not overly concerned, as she noted that going forward, the District would focus intently on the teacher’s instruction and use other methods allowed by school code, such as neglect or incompetence, to pressure the teacher into resigning.

Despite some concerns, the TEM was perceived to provide meaningful value or worth in District C. All three of the factors raised by Seashore-Louis & Dentler (1998) were evidenced. The topic of supervision and evaluation was important, based on the personal belief of the key decision maker associated with TEM interpretation. It was useful in that its approach to supervision and evaluation provided a lot of potential opportunities on which previous models
had not been able to capitalize. The TEM provided gains in administrator knowledge about instruction and to a lesser degree administrator’s ability to detail specific concerns with under-performing teachers. It is important to note that much of the perceived benefit represented in these factors stems back to the personal belief and the passion of the Director of C&I for a tightly coupled supervision and evaluation plan. Because of her role in the district, as the key individual designing TEM implementation, and the alignment of the TEM with her personal beliefs, the model gained the perception of value or worth outlined above. Had the Director little interest in supervision and evaluation, the importance factor might have been greatly reduced, and as such, the usefulness and benefits would have been much less important.

**Competing Factors.** There were a number of other factors taking time and resources in District C; however, because of effective “siloing” by the Superintendent and strategic decision making by the Director of C&I, the TEM did not compete with nor interfere in the districts direction. The Superintendent described three major initiatives of the district as integrating technology into instruction, increasing financial efficiency, and implementing the district’s comprehensive plan. The technology emphasis was marked in its priority to the district, its cost, and the time it took for professional development. At the high school, all students had their own computer, and at the middle school there was one computer for every two students. Teachers were expected to use technology for instruction, and the overwhelming amount of the four days of staff development built into the yearly contract was dedicated to this topic. The Superintendent stated, “It is our goal to have technology become a part of everything they [teachers] do with the students.”

The focus on efficiency was driven by budgetary concerns and aimed to get the most out of the resources at hand and limit unnecessary spending. The process of becoming more
efficient involved working with all stakeholders, including teaching staff, and cost time for meetings and discussions as well as put a limit on how much money could be spent on any given initiative. The comprehensive plan outlined the values of the district and covered student achievement, community engagement, and financial stewardship. While the TEM fell under student achievement, the Superintendent was clear to state that supervision and evaluation was just one of many things that were important, and could not be given undue weight — “The teacher effectiveness model is important, but the district is much bigger than teacher effectiveness.”

All of these factors could have suffered had the TEM been perceived as requiring priority status in terms of time or cost, or, the TEM could have been largely buffered from the district given the importance of these local factors. However, the Superintendent worked hard to delegate different responsibilities across his central office staff, allowing his Director of C&I to focus on the TEM while he and his Assistant Superintendent focused on the rest. The Superintendent spent much of his time building and maintaining relationships with the community (community engagement in the comprehensive plan) as well as communicating the budgetary needs of the district (financial stewardship from the comprehensive plan). The Director of C&I, understanding the time and cost that would come with the TEM, proactively applied for and won Race to the Top grant monies. These funds allowed the District to pay for substitutes and train teachers on school days, instead of competing with technology for the four contracted in-service days.

District C had many factors competing with the TEM. Evidence suggested that the main reasons why these factors did not interfere with the model had to do with how the TEM was assigned to one key central office figure (while others took on the rest of the districts initiatives) and because that individual had the personal beliefs as well as the professional experience to
know how to bring a new initiative into an already busy district. This was hallmarked by the Director of C&I’s decision to apply for and ultimately win federal grant money to help pay for professional development related to the TEM. As such, the TEM could function beside the other district factors, instead of actively competing with them.

**Social Interactions.** Central office staff were part of social groups inside and outside of the district. Evidence did not support the idea that social networks outside of the district influenced how the TEM was implemented; however, evidence did suggest that internal social groups made a difference. The Superintendent used his peers to understand different aspects of the TEM, but did not make changes to District C’s model based on his conversations with other districts. Similarly, the Director of C&I noted the position of others regarding certain components of the model (such as SLOs) but did not seem to have changed her opinion based on the thoughts of others. In an example from early in the process and contrasting District C to others, she noted, others “were not taking Race to the Top money because they were not going to accept money and be bribed into the system. There was no reason to change their current systems. I was the one to say the old system was antiquated, I don’t care what model you’re using, you’re not doing it well, and it’s going to be legislated.” This example showed how central office staff in District C were aware of what others were doing yet had personal beliefs so strong as to not be swayed by social pressure.

Both the Superintendent and Director of C&I mentioned conversations with the teacher’s association. The Director noted that the union president was one of the teachers involved in the TEM pilot and that she (the president) was on the committee involved in TEM implementation. The Director reported, “She’s involved, everybody feels there is credibility. If she has any concerns she comes to us.” Because of this communication and networking within the district,
certain details such as exactly how SLOs looked or how IEP growth was calculated were worked out peacefully.

Overall, the social influence on TEM interpretation was limited. It seemed to be overshadowed by very strong personal beliefs on the part of the Director of C&I and on the fact that the TEM was not interfering with any of the other district initiatives (and thus, not causing problems at the Superintendent’s level).

**Interaction of Factors in District C.** The interpretation of the TEM in District C is a story that hinges on the personal beliefs of the Director of C&I. Firestone & Shipps (2005) documented the role personal beliefs can play when interpreting reform related information. Consistent with this, the alignment between what the Director of C&I believed and how she perceived the TEM to fit those beliefs led to an intended interpretation consistent with the original purpose of the reform (what Spillane, 2000, called second order change). The Director of C&I had a long standing passion for supervision and evaluation and a long standing concern for the ineffectiveness of models in practice prior to the TEM. The TEM itself presented to her as a tightly coupled system that had the potential to make drastic improvements in areas where she had noted concerns historically. Paired with her influential role in the district, this personal belief naturally linked to assigning value or worth to the TEM and to realizing a usefulness and potential benefit from its implementation.

Past experiences (Spillane, 2000) and knowledge of reform (Coburn, 2005) are two additional factors that played into the interpretation of the TEM in District C. The Superintendent had learned over time how to separate, or silo, certain tasks among his central office staff; and did so with the TEM so as to allow one person to focus on it while others could focus on the other important initiatives and issues in the district. The knowledge of past reform
efforts as well as the components (and complexity) of the TEM allowed the Director of C&I to prepare for the great number of other tasks that required attention, time, and funding; including strategically applied for federal funds that allowed the TEM to move forward without interfering with local priorities. Despite strong social networks, the Superintendent or the Director of C&I did not change their interpretation of the TEM based on conversations with others.

**District D**

**Background.** District D was in a rural distant locale (National Center for Educational Statistics, 2015). The district was 194 square miles in size with 2,651 students attending five schools. The school performance profile (SPP) scores in these schools ranged from 72.5 to 84.8. Thirty four percent of the student population was classified as economically disadvantaged. The annual budget for District D was about $42 million, and data from the Department of Education suggested there was a fund balance of about $6.5 million (PDE, 2015).

The Assistant Superintendent and Superintendent were interviewed for the study. The Assistant Superintendent had been in the central office for seven years working with curriculum and, at one time, technology. Prior to this, he was an elementary principal in District D for two and half years. He had also worked as an assistant principal and classroom teacher outside of the district. The Superintendent was in his 12th year in that position. Previously, he had worked in other districts as an assistant to the superintendent, principal, athletic director, and classroom teacher. The Superintendent noted that he was both a resident and graduate of District D.

**Past Experiences & Personal Beliefs.** Firestone and Shipps (2005) documented the role personal beliefs, norms, values, and experiences play in how reform information is interpreted. In District D, the Superintendent’s background and experience played a meaningful role in his interpretation of outside information. Interestingly, though, and unlike that suggested by Honig
(2003), Spillane (2000), or Spillane, Reiser, and Gomez (2006), it was not the Superintendent’s experience with specific reform components that led to a second order, or function focused, change; it was his experience with reform and outside mandates in general that led to first order, or form only, type of change. The Superintendent had served as the district leader for 12 years and had spent almost three decades working in public education, across a number of districts and two different states. In addition, he was active in state and federal education leadership associations and worked on committees and focus groups to help shape policy. As a result, he had learned the many factors that influenced state mandates, as well as what those mandates meant, and often what they did not mean, at the local level.

The Superintendent used his experience to guide the lens through which District D made sense of the Teacher Effectiveness Model (TEM). He explained the purpose of the model as a way for the State to meet federal NCLB waiver requirements, noting, “At the end of the day, it truly was political, mandated, if you want relief from NCLB.” As such, he viewed the mandate not as a benevolent attempt to improve teaching and learning, but more so a bureaucratic necessity. This perspective led the Superintendent to take a stand both in and out of his district, telling colleagues, “We cannot let this define us.” By this he meant that District D (or any other district) should not make massive changes nor get overwhelmed by the management of the TEM process. As a result, the Superintendent led the district to buffer a great deal of the district’s take on the TEM. The Superintendent’s experience allowed him to see the TEM as part of what his district was already doing. He noted, “When it came out, I told our staff not much was going to change, don’t get caught up. If I have a good teacher, the last thing I want them to worry about is whether they are distinguished.” He noted, “The TEM for District D was the strongest mandate with the least effect. We already used Danielson and we just changed what we used to
call teacher goals to SLOs.” The Superintendent’s experience also taught him, however, that he needed to be measured in his comments. When discussing with whom he shared his opinions, he noted, “It’s not going to be something public.” Finally, he had learned that there was opportunity in most reform efforts, regardless of the drawbacks. As such, he worked internally with district office staff to identify parts of the TEM that would benefit District D and made sure the district took advantage of those. Talking about the SLO process, he noted, “The SLOs fit nicely, talk about growth, strategic planning, and goal setting.”

Central office staff in District D expressed strong personal beliefs in regard to the purpose of supervision and the divide that should exist between supervision and evaluation. In addition, a strong belief in serving the local community and local needs emerged from the data. These beliefs were in contradiction to the purpose and underlying philosophy of the TEM, and when factored in with the Superintendents experience, sent a strong push through District D to implement only as much of the TEM as was necessary.

The concept of growth was embraced as the real purpose of supervision. The Superintendent said, “Teacher growth and preparing kids for the next step, it’s not a ‘gotcha,’ it’s growth, plain and simple.” The Assistant Superintendent thought supervision should be about teachers getting better, and noted that supervision works best when there is no “threat” of evaluation. By this, he was referencing a safe environment wherein teachers would take risks and push themselves beyond their own comfort zones. He explained, “Supervision needs to be as free of threat as possible. I want the veteran teacher taking a risk with technology or a new teacher objectively identifying her needs.” Both administrators worried that this risk would not be taken, and growth would be limited, when evaluation was tied so closely to supervision, as they both saw the Teacher Effectiveness Model (TEM) doing. “I struggle with the new model
because supervision and evaluation have been married so closely together” noted the Assistant Superintendent.

Local control and doing things that made sense in the local context were strong beliefs in District D. Both the Superintendent and the Assistant Superintendent used the term “District D-ize” to reference how any information from the outside was changed so as to fit the local context. The Assistant Superintendent explained the school board’s expectation for the administration to look closely at any outside initiative such as the TEM, the SPP, or the old AYP model, and consider how it fits the local context, not just accept it as presented. He noted, “The board expects us to test the boundaries of state mandates to make sure it fits the local community, not break the rules but make sure it fits for District D.”

Cumulatively, the experience of the Superintendent and the personal beliefs of the Superintendent and Assistant Superintendent played an overwhelming role in how the TEM was interpreted in District D. The district was not in the habit of jumping on any new reform as presented, and always worked to fit outside information into the existing local context. Moreover, the philosophy of the TEM, as perceived by central office staff, was in contradiction to local beliefs regarding supervision. As a result, much of the TEM was interpreted in a way consistent with Spillane’s (2000) idea of first order change (or Honig’s (2003) idea of form focused change) wherein some effort was made to look as if things were being done, but ultimately, the district continued to work through supervision as it had been doing prior to the TEM.

**Perceptions of Value or Worth.** District D appeared to derive limited benefit from the TEM in that it provided some leverage and additional focus on areas thought to be important at the local level. In large part, however, the TEM did not present as useful, mostly due to the
perceived value or worth the prior model of supervision and evaluation was given and the contradiction in philosophy between the quantitative nature of the TEM and the personal beliefs of central office staff toward a more pure growth approach to supervision. The old model was also noted to deal with under-performing teachers in a much more efficient manner than the TEM.

District D held strong beliefs in regard to the growth nature of supervision and had established a model that reflected this belief long before teacher effectiveness was mandated. The old model was described as having three options through which staff would rotate over three years. One of those three years would be a clinical observation model wherein teachers would be formally observed and would provide feedback on areas of strength and need. Interestingly, District D used the Danielson model for their clinical observation framework. The other two options were self-directed or group action research. The Assistant Superintendent described the focus of all three options as “skills and growth. Because there was no rating involved, staff could focus on making improvements to practice that directly affected instruction.” The Superintendent described how his administrative professional development prior to the TEM included a number of days with an outside expert focused on how to use the Danielson model to get teachers to reflect upon their instruction and what they could do to improve it. Contrasting the old model with the TEM, the Superintendent noted, “Scoring is the big difference. With the old model, there was no scoring during the course of the year. The focus was on reflection.” The Assistant Superintendent said, “Our model was as absent evaluation as it could be. Conversations were much more about learning than they were meeting a certain benchmark.” This point was important to central office staff. The old model had been created locally and purposefully designed to be qualitative. The TEM, by adding a quantitative interpretation to
supervision through rubric-ratings on Danielson domains and components, was perceived to take away from the growth focus of the old model and shift teachers attention more toward the category in which they ultimately would fall (proficient, distinguished, etc.). In this way, the TEM actually took away the usefulness and benefits that had already been realized through the old model. Thus, the shift to TEM reinforced the first order or form only focus interpretation of the TEM.

The TEM did not provide District D with any advantage in working with underperforming teachers. In the model that existed before the TEM, teachers would be identified as under-performing through clinical supervision or other walk-through observations. Once identified, additional observations (up to eight) would take place, and if warranted, a focused supervision plan would be developed. The focused plan would detail a few specific areas that would be targeted for growth and various types of support that might lead to growth were included. The team that would develop the plan included the teacher, building and central office administration, and a representative from the teacher association. The Assistant Superintendent described the process as intended to “coach up, not coach out. It was not to catch someone doing something wrong, but to help them do things correctly.” The Superintendent noted that the old process took “an inordinate amount of time” and included a lot of observations and meetings. However, the process was effective in growing many teachers to a proficient level or demonstrating to individual teachers that it might be in their best interest to move on from the district and perhaps education in general. In the first few years of the old model, seven to eight teachers were on focused supervision plans each year. Eventually, there were only two or three teachers each year, and the Superintendent thought this reduction was due to how effectively the
model either grew teachers to proficient levels or sorted out those who were not going to be good teachers.

The TEM was viewed by central office staff as detrimental in terms of working with under-performing teachers. While there were a limited number of examples, those that existed were profound and clearly violated the local culture and expectations of District D. The TEM did not prohibit the use of a focused supervision plan, and thus, the district continued to use its old model. However, instead of being able to focus on just a few targeted areas, the mandate to use, and how to use, the Danielson approach, with its four domains and many more components, increased the breadth of evaluation. The increase in areas paired with the addition of numeric ratings to the model (and the idea that few teachers would earn failing ratings in all areas) meant that overall observation ratings were almost never low enough to justify an unsatisfactory conclusion. As well, the addition of the data component through the SPP and SLOs seemed to inflate an under-performing teacher’s marks and increase the difficulty in justifying an unsatisfactory rating. Central office staff told about a teacher who was underperforming and subject to a focused supervision plan. This teacher had been observed no less than 56 times, compare to eight in the old model, over the course of the year in an attempt to gather adequate information. The Assistant Superintendent noted, “The table is tilted toward the teacher to be satisfactory [rather] than unsatisfactory. Prior, this would not have taken as much time, it would have been more qualitative, [and] now it’s just a numbers game, and dismissal is nearly impossible.” At the end of the year, the teacher’s evaluation was rated as unsatisfactory, despite the Superintendent’s concern that such a rating would not stand up to challenge. At the time of interview, the case was in arbitration and the district was unsure as to whether or not they would be able to remove the teacher from classroom. Frustrated, the Superintendent asked, “Would
you want your kid in that classroom? That’s my biggest struggle as a superintendent, the way the State went about this model has kept that teacher in the classroom.” Lamenting what was lost because of the TEM mandate, he noted, “If I have a system that is objective and uses the same observation model and someone’s not making growth, I can give them due process, but if they don’t respond I can give them an unsatisfactory. With the [TEM] I have to rate that person satisfactory, the TEM is a safety net for those who should be unsatisfactory. That galls me.”

Despite its many frustrations, the TEM was not a total loss in District D. The Danielson framework had been used in the district for some time, but the mandate for it increased the attention paid to the full breadth of each domain and the respective components. The shift to a quantitative rating increased the need for teachers to gather observable evidence in each component, as well. The Assistant Superintendent noted the latter point changed some from “administrator versus teacher to now being based on evidence, which has been a benefit.” Another small, but very helpful, result of the TEM was alignment within buildings among teachers who were on formal observation or self-directed professional supervision models. Under the old model, there was no organization as to when someone would do what. As a result, it could be difficult for a group of third grade teachers to all work on the same project as a group goal, as one could be in the formal observation part of the cycle. With TEM implementation, the district had to organize cycles and did so strategically so that common teams would either be in formal observation or self-directed plans at the same time. The focus of the SLOs was also helpful. The district had always created building level goals and encouraged individual teachers to focus their own goals and growth toward such; however, the district was able to leverage the SLOs in a way that forced all teachers to embrace building level goals at the individual level.
The Superintendent believed the alignment from teacher level to building level resulted in a benefit to the district.

Special note should be made in regard to how District D bridged and buffered parts of the TEM. This was a purposeful and often explicit process, with central office staff stating in no uncertain terms that outside information was always “District D-ized” when it was brought in. The model as a whole was not thought to be needed in District D, as there was already a supervision and evaluation plan that seemed to be reaching the local goals of teacher reflection and growth and was allowing the district to either grow or dismiss under-performing teachers.

As a result, the Superintendent’s general response was to buffer any part of the model that did not align with district processes or goals. To this end, he noted, “We provide the score sheets that are asked for [to PDE] but they absolutely do not guide the way we move as a district in regard to supervision.” The Danielson model was changed to add the quantitative ratings, but this was minimized in the principal-teacher interaction throughout the year, with a continued focus only on reflection and growth. The Superintendent said, “That’s ours and it’s what we do.” The district only emphasized ratings at the end of the year, “The evaluation is the game the State gave us, and we are playing that game.” Other parts of the model simply were not implemented at all. In example, there were three components of the data portion, the SPP, teacher specific data, and elective data. District D completely buffered the teacher specific portion and simply used the elective data to cover the rest of the data portion of the model.

Talking about this and a very specific portion of teacher specific data related to special education, the Assistant Superintendent noted, “We have worked for a long time to have all staff take ownership of all students. Our fear was that if IEP growth was in the mix, staff would not
want special education students in their classrooms. We worked so hard to overcome that we didn’t want to revert.”

Other parts of the model were bridged to existing initiatives. The SLO process was used in a way that individual teacher goals, in the form of SLOs, were linked directly to more broad building level goals. At the middle school, there had been a shift toward students completing “cornerstone tasks” at the end of the year. The district incorporated this requirement in a way that teacher SLOs were in some way related to student performance on these tasks, and as a result, allowed teachers to feel accountable for student performance directly related to the building goal. Other SLOs were related to the district’s 1:1 technology initiative or student research projects related to bee extinction or the use of fertilizer and its impact on the environment, both pressing issues in the local community.

Seashore-Louis and Dentler (1988) wrote about the degree to which reform is important, useful, or will lead to a gain in a specific district; and how such influences interpretation. In District D, while there were some benefits and potential gains, the overwhelming majority of the new TEM took away from the benefits and gains found under the already existing model of supervision and evaluation. As a result, much of the TEM was buffered (Honig 2003), leaving a form focused, first order (Honig, 2003; Spillane, 2000) change in process only. Central office staff were clear and comfortable on the approach they had taken, suggesting that their experiences and personal beliefs compounded with the (lack of) perceived value to create a sense that it was not in the district’s best interest to make substantial changes to the already existing system.

**Competing Factors.** There were many other things going on in District D when the TEM was introduced, and despite the overwhelming efforts to minimalize the impact of the TEM
on the day to day practices of supervision, implementation consumed a great deal of time, especially for the Assistant Superintendent. Other local initiatives suffered as a result, and while frustrating, this did not appear to play a meaningful role in how the TEM was interpreted. Factors such as the roll out from the state were also frustrating, but did not seem to play into the local interpretation to the degree that experience, beliefs, and lack of perceived value or worth did (as described above).

The Assistant Superintendent, who was charged with interpreting and designing the implementation, had to put many projects on hold. He described the things that were most likely to get less attention as those for which there was not as much accountability. For example, the district had started to shift its libraries to “learning commons.” This was a local decision only and as a result of the TEM, he noted, “The learning commons are not where I want them to be right now. I just didn’t have the time.” Other projects carried on, but did not have the Assistant Superintendent involved. “I removed myself from the professional development committee and had to have the Director of Curriculum and Instruction take over.” Despite the role TEM played in delaying local initiatives, the Assistant Superintendent was not overly frustrated. He noted, “I have to think there is a reason why the [TEM] was mandated and even with its problems, we need to implement to some degree.”

The TEM was described as different from other reforms the district and its central office staff had known in the past. The Superintendent noted, “The roll out was horrific.” He cited a “lack of a road map” for the total model and pointed to things like the release of the SPP in October, when teacher evaluations are completed in June, as evidence that the model was not presented very well at all. Additionally, the conversion of the Danielson model from a qualitative approach focused on reflection and growth to a quantitative approach focused on
evaluation was poorly conceived at the State level. The SPP, which made up 15% of a teacher’s evaluation, was originally supposed to be comprised of multiple measures, but the Superintendent noted that it was mostly made up of PSSA and/or Keystone test results. Thus, achievement level or growth was based only on achievement level or growth based only on the high stakes snap shot. While clearly frustrating, the roll out did not appear to play a meaningful role in how the TEM was interpreted.

**Social Interactions.** Social networks, both in and outside of the district, helped to influence the interpretation of the TEM in District D. Outside of the district, social networks helped central office staff develop a common interpretation of the TEM in general as well as understand different TEM components (Seashore, Febey, & Schroeder, 2005). Inside the district, similar to that described by Searshore-Louis and Dentler (1988) and Spillane et al. (2002), the specifics of TEM were discussed, refined, and interpreted in terms of how they could be used locally.

Conversations among the Superintendent and his colleagues in other districts focused on the overall embrace of the TEM model. The Superintendent noted that there was a great deal of difference in interpretation among his group of peers. He noted, “For some, it was the sky is falling, but for others, we didn’t get too bothered by it.” The Superintendent of District D took the latter approach, and this appeared to be influenced by his experience and background. He reported talking to other superintendents about “the state having no system to audit TEM” and as such, “As superintendents we can focus on what works best in our districts and change the model to align to our needs.” In this way, the social network may have helped District D to be emboldened to buffer much of the model, knowing that at least some other districts were taking the same approach.
Conversations among the Assistant Superintendent and his peers centered on how to figure out specific components of the TEM model and were vital in figuring out what the TEM should mean in practice. He noted, “My learning was much greater with other district administrators than it was through the state or any formal training.” He described the typical process wherein his colleagues and he would attend a training with the State, then gather as a group and discuss “how to figure out what it was the State had to say. This was much more meaningful and relevant.” Offering an example, he explained how teacher-specific and elective data were explained and adjusted from the State level. This became very confusing and potentially time consuming, and would provide minimal change to any given teacher’s overall rating. As a group, the assistant superintendents and directors of curriculum talked about different ways to implement these two pieces at their own local level.

The group information was very helpful; however, as the Assistant Superintendent noted, “Our districts are all different, and we have different cultures, so what one district can do another cannot, depending on local things like the union, community values, and the school board.” Describing how students were attributed to teachers in the valued added model (to measure teacher effect), District D listened to but did not go in the same direction as other districts. “They would attribute PSSA results to a reading specialist working with 3rd graders, we didn’t think a reading specialist was focusing on 3rd grade reading standards when intervening in a case like that, so we didn’t go along with that.” When considering the SLO process, other districts were very concrete in their design, leaving little room for teacher input. The Assistant Superintendent noted, “This made but sense, but absolutely was not the direction we wanted to go in. In our minds, the SLO should be evidence of professional growth linking to student growth or achievement. So, we specifically said we’re not going in that direction.”
This latter point led to the social interaction within District D. Both the Superintendent and Assistant Superintendent learned with and from their colleagues outside of the district and brought that knowledge back for discussion. Conversations with each other as well as central office staff in curriculum and pupil services allowed for local interpretations that were consistent with local beliefs. The Assistant Superintendent reported, “We looked at what we had to do, then looked at what we could do to make it work here. We knew our prior model was working for us, so we’d come back from a training and debrief, try to figure out how to make it work within our older model.” The Superintendent noted, “We looked at how we could manage the model without losing our own [prior] model. How could we get the TEM into our [already existing] building goals?” He went on, “We also worked to make sure we didn’t get caught up in the management of it. We didn’t want our administrative team to be tripped up by a cumbersome process.” The Superintendent offered an example of how his district team collaboratively had to respond and work out a way to deal with the management burdens brought about by the TEM. The high school principal had recently left the district, and the Superintendent realized that the overwhelming amount of paperwork associated with the TEM probably played a role in his departure. To ensure this did not happen to anyone else, the district office staff brainstormed ways to reduce the work load for principals, including using department chairs to help with observations.

Social interaction led to higher levels of learning in regard to how to interpret the TEM, and evidence of this process was clear from the reports of central office staff in District D. Many have documented the advantages of social capacity in terms of learning new information (Coburn, 2001; Sanders, 2012; Seashore-Louis & Dentler, 1988; Spillane et al., 2002) and the capacity found in and around District D was marked. As a result, and when considered in the
broader context of factors such as experience, personal beliefs, and the lack of perceived value or worth of the TEM, central office staff were able to better understand what they could buffer and where there the TEM offered leverage enough to bridge.

**Interaction of Factors in District D.** District D provided an example of how a number of different factors can align to drive the direction of interpretation. The background experience of the Superintendent was such that he had learned over a long career to have trepidation when it came to politically motivated change. He had also learned that districts were sometimes better suited to pause and consider continuing their local practices when this was in their best interest. The personal beliefs of central office staff were in contradiction to the philosophy of the TEM, especially in the focus on teacher growth versus teacher ratings. When paired with the Superintendent’s knowledge, this belief set the stage for more buffering of TEM concepts and a focus on first order change rather than a cultural shift in the fundamentals of supervision. This perspective was reinforced by the lack of perceived value or worth the TEM brought to District D in that it limited teacher growth by changing the focus to a quantitative score and made working with under-performing teachers much more difficult. Finally, the social capacity of central office staff, both in and outside the district, was strong, and it helped the Superintendent and Assistant Superintendent to understand various components of the TEM as well as embolden much of the buffering that took place.

**Cross Case Analysis**

The cases studied offer a view into four school districts and how each made sense of the Teacher Effectiveness Model (TEM) reform. While the analysis above provides a detailed look within the individual districts, consideration of similarities and differences across the cases allows for a deeper understanding of the sensemaking phenomenon and how it was influenced by
Each of the factors considered (Stake, 2006; Yin, 2003). Each of the factors played a role in all of the cases; however, factors related to perceptions of worth or value, personal beliefs and social interactions were marked as far as their contribution to how interpretation was made in each district. In fact, it was the complex interaction between these three factors that seemed to account for the majority of reason as to why a district embraced (bridged) certain parts of the TEM and avoided (buffered) others.

Table 1: District Demographics

<table>
<thead>
<tr>
<th></th>
<th>District A</th>
<th>District B</th>
<th>District C</th>
<th>District D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enrollment</td>
<td>3821</td>
<td>3140</td>
<td>4956</td>
<td>2651</td>
</tr>
<tr>
<td>Square Miles</td>
<td>16</td>
<td>84</td>
<td>76</td>
<td>194</td>
</tr>
<tr>
<td>Percent Economically Disadvantaged</td>
<td>29%</td>
<td>20%</td>
<td>34%</td>
<td>34%</td>
</tr>
<tr>
<td>Locale</td>
<td>Suburb: Large</td>
<td>Rural: Fringe</td>
<td>City: Small</td>
<td>Rural: Distant</td>
</tr>
<tr>
<td>Budget</td>
<td>$55 Million</td>
<td>$48 Million</td>
<td>$70 Million</td>
<td>$42 Million</td>
</tr>
<tr>
<td>Fund Balance</td>
<td>$18 Million</td>
<td>$8 Million</td>
<td>$18.3 Million</td>
<td>$6.5 Million</td>
</tr>
<tr>
<td>Number of Schools</td>
<td>8</td>
<td>6</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>SPP Range</td>
<td>76.6 – 90.8</td>
<td>72.8 – 88.2</td>
<td>75.8 – 94.4</td>
<td>72.5 – 84.8</td>
</tr>
</tbody>
</table>

Demographic information for each district is presented in Table 1. The four case districts studied included rural, suburban, and city locales. The smallest district (District D) had 2,651 students while the largest (District C) had 4,956. District B had the least amount of economically disadvantaged students (20%) while District C and D had the most (34% each). All four districts had adequate fund balances relative to their overall budget. School Performance Profile (SPP) scores were similar in that no school in any of the four districts was performing below a 70. Fund balance and SPP data suggested that none of the four schools studied were at great risk for the limited capacity that has been linked to resistance to change (Ogawa, Sandholtz, Martinez-Flores, & Scribner, 2003; Olsen & Sexton, 2009). Finding districts like these was important given the argument articulated by Spillane, Reiser, and Gomez.
(2006) that local interpretations of reform are often due to factors related to background knowledge or social learning, not just lack of capacity or resource.

**Perceptions of Value or Worth.** Seashore-Louis and Dentler (1998) proposed that the degree to which a reform was perceived to be important in a district, the degree to which a reform was perceived to be useful, and the degree to which a reform was perceived to provide a benefit for the district influenced how sense would be made of any new initiative. These three value- or worth-related factors played a marked role in how the Teacher Effectiveness Model (TEM) was interpreted in all four districts. Evidence suggested that the topic of supervision and evaluation was important in all four cases and as such, (consistent with Seashore-Louis and Dentler, 1998), the TEM received a great deal of attention in all of the cases studied.

Considering usefulness and gain and looking at the model as a whole, Districts A and D already had supervision and evaluation procedures in place with which they were very happy. Both districts expressed a degree of frustration at the fact that an outside entity (the State Legislature through the Department of Education) was meddling in local affairs when they already were doing quite well with the topic at hand. In contrast, Districts B and C did not have strong supervision and evaluation models in place, recognized this deficiency, and saw the TEM as a way to improve. The response to TEM in Districts B and C was much different than in A and D. In Districts B and C, the TEM was pushed toward implementation with the intent to embrace many of its subcomponents. In Districts A and D, implementation was limited to legal requirements. The difference between the districts was based on whether or not the initiative was perceived to address an area of need or not, and the benefit-related factors of use and gain played a meaningful role in explaining the difference.
The worth associated with TEM in each of the districts was evident in the observation, student learning objectives, and teacher-specific data subcomponents. Three of the four districts expressed a sense of usefulness and benefit from the Danielson model. District D had been using the Danielson framework prior to the TEM, but with the new model was forced to change from a qualitative use to a more quantitative ratings-based form. Other than this example, the other three districts found that Danielson gave them a framework through which their administrators and teachers could understand the important components of teaching, and this led all involved to focus more on observable evidence rather than simply on opinion.

All four districts found some benefit in using the TEM as a state mandate to leverage local initiatives related to supervision and evaluation. Interpreting outside information in this way has been described by Honig and Hatch (2003) as “bridging” and this process was exemplified many times over in each of the four districts. The Student Learning Objectives (SLOs) stand out as a primary example as they were used in each case study site to help increase teacher focus on student outcomes and support local efforts that had been ongoing but had not yet taken hold.

All four districts struggled to find value or worth in the teacher specific data subcomponent. Implementation of this subcomponent was marked by the perception of mixed and confusing messages from the Department of Education to local districts, as well as one prominent last minute change related to including the growth of students in special education. Staff from all districts studied expressed frustration with what appeared to be a lack of clarity over what this subcomponent should represent. To staff, the subcomponent was not useful in supervision or evaluation and provided no gain or benefit. The time and resource needed to complete teacher specific data was perceived to be much greater than the worth of the outcome,
and this led most central office staff to either partially or totally limit implementation of this section of the model. This limiting process of selectively ignoring parts of a reform that are perceived as being less efficacious is labeled “buffering” (Honig and Hatch, 2003).

Looking beyond subcomponents, there was one additional theme that stood out as it related to the perception of value or worth of the model. Across all four cases, the TEM initiative was noted to be ineffective in identifying underperforming teachers. As a result, its usefulness was questionable when administrators were working toward termination. Each central office administrator interviewed gave examples of under-performing teachers and the difficulty the TEM presented when considering dismissal. Referencing their beliefs in removing ineffective teachers from classrooms, all four districts reported to use other means to work toward dismissal, often going outside of the evaluation model to methods listed in School Code related to incompetence or neglect.

**Personal Beliefs & Intent.** Firestone and Shipps (2005) noted the potential influence of personal beliefs, norms, and expectations on how sense is made of outside or reform-oriented information. In the current study, personal beliefs interacted with perceived value or benefit, leading central office staff to judge what was useful and of benefit in the TEM based on their beliefs, and these two factors (perception of value or worth and personal beliefs) appeared to contribute more to how sense was made of the TEM than did any other individual or set of factors.

The interaction between personal beliefs and the perceived value or worth yielded by the TEM was exemplified in the differences between Districts B and C versus A and D. In the latter two districts, a successful model of supervision and evaluation existed prior to the introduction of the TEM. The concept of supervision and evaluation was important to staff in all four
districts, but in Districts A and D personal beliefs related to supervision and evaluation had been satisfied. It was only in Districts B and C where a discrepancy between beliefs and practice existed, and the TEM was perceived to have more potential worth or benefit. These two districts had attempted to put models in place prior to TEM, but each had not met with the success desired. Through the TEM, they saw the opportunity to leverage practices aligned with their beliefs about supervision and evaluation. As such, Districts B and C looked to bridge many more parts of the TEM than did Districts A and D.

As might be expected, the components of the TEM that aligned with the beliefs of the central office or the culture of the district were embraced (bridged) while those components that contradicted personal beliefs or local norms were mostly buffered. Central office staff in all four cases listed a personal belief that teacher supervision should be about professional growth. In Districts A, B, and C, this belief aligned with the approach of the Danielson model, wherein the four domains and numerous components allowed observers and teachers to reflect on specific areas of strength and need. Administrators designed teacher goals linked directly to factors that would allow for improved instruction. As such, this part of the TEM was embraced in these three districts. The Danielson model was already being used in District D prior to the TEM, and the changes of the TEM were perceived to actually take away from the potential for growth and reflection. In this way, the TEM contradicted personal beliefs and the use of the Danielson model as prescribed by the state was buffered to the greatest extent possible.

**Competing Factors.** Firestone and Shipps (2005) noted the role other sources of accountability in a district play when it comes to interpreting new ideas or reform. These authors suggested that the more going on in a district the harder it was to focus on new information and/or align it with pre-existing initiatives. There was very little evidence to support the idea
that the TEM was limited by competing factors in any of the four cases; however, the sample of
districts was such that meaningful capacity-related factors (financial hardship and poverty) did
not exist.

All four districts had a great number of competing factors but responded in ways that
allowed the TEM to receive the attention and resources that were needed. In District A,
additional staff were hired to keep other local initiatives moving forward. In Districts B and C,
the Superintendents strategically “siload” central office staff duties so that a single administrator
could focus on the TEM while others focused on other necessary work. District D was similar in
that other initiatives were delegated or in some cases simply put on hold. All four districts
worked to bridge as much as they could to the TEM. Each appeared to see the TEM as an
opportunity to leverage local initiatives and as such, districts shifted some academic factors from
“competing” to simply being aligned within the TEM process. This was also exemplified across
cases with the student learning objective (SLO) process. Instead of the SLOs competing with
other local pushes, each district bridged the SLOs to already existing (or desired to be existing by
the central office) practices related to student or building growth. In this way, practices related
to growth did not compete with the SLO, but rather, became one and the same.

Past Experiences & Reform Knowledge. The literature suggests that the more that is
known about a reform and the larger the amount of experiences with reform, the greater the
likelihood a reform will be implemented as intended (Coburn, 2005; Spillane, 2000; Spillane &
Thompson, 1997). This is what Spillane (2000) and Honig (2003) refer to as second order or
function-focused change. In the four cases examined, those central office staff charged with
implementation all had solid understandings of the concepts underlying the components of the
TEM, either before or gained during implementation. However, this did not seem to be as
closely related to second order change as it did effective bridging (through leveraging). Specifically, because central office staff appeared to have a basic understandings of concepts like reflection, key to the Danielson model, or the role of formative and summative assessment in understanding student learning, a key to the SLO process, they were able to use the TEM to move forward ideas that were important to the district or aligned with their personal beliefs. This finding differs from the idea that knowledge of reform leads to second order type change (Coburn, 2005; Spillane, 2000; Spillane & Thompson, 1997). Findings here suggest that knowledge of reform appeared to lead to a deviation from the first order versus second order continuum, and instead, contributed to how effectively outside information could be used to leverage pre-existing local initiatives.

To the latter point, it is worth noting that many of these pre-existing local initiatives were aligned with the stated purpose of the original federal waiver, the policy that drove TEM. In 2010, the USDoE wrote that the waiver was intended to “increase the quality of instruction for students and improve student academic achievement in the State and in LEAs” (p. 5). As such, it could be concluded that, consistent with Coburn (2005), Spillane (2000), and Spillane & Thompson (1997), knowledge of reform would lead neatly to second order type change. However, those interviewed for the study clearly perceived the TEM to be driven by other reasons such as to increase the number of teachers dismissed. Central office staff thought of their own local initiatives as separate and distinctly different from anything related to the TEM or policy surrounding it. Thus, findings are interpreted to differ from that suggested by Coburn (2005), Spillane (2000), and Spillane & Thompson (1997), and instead suggest that background knowledge led more toward effective leveraging of pre-existing local initiative’s instead of an embrace of the spirit and values behind the TEM.
The past experiences of superintendents allowed each to keep the TEM initiative in perspective, although this was more pronounced in some districts than in others. The Superintendent in District D had been in his position for 12 years, and his experience had taught him that even state-mandated change did not need to mean all that much to the local level. As such, in a district that already had a strong supervision and evaluation model, his central office staff buffered a great deal of the TEM. Similarly, the Superintendent of District C had been in the central office for 15 years, six of which in his current position, and he took the position that while the TEM was important, especially in his district that was in need of an improved supervision and evaluation model, it was just part of the much bigger district picture.

**Social Interactions.** The role of social interactions in learning is well documented. Crossan, Lane, and White (1999) provided a general framework detailing how learning shifts from the individual to the group. Spillane and Thompson (1997) wrote about how effective social capacity leads to the spread of reform principles across an organization. Both Seashore-Louis and Dentler (1998) and Spillane et al., (2002) detailed how new ideas were refined and disseminated through social learning. Cumulatively, the social component of learning is where the organization makes sense new information, figuring out what it means and how it can be applied. Evidence from the four cases found high levels of social capacity across the districts and a multi-level approach to interpretation.

At the first level, staff responsible for the implementation of the TEM (usually an Assistant Superintendent and/or Director of C&I) would meet across districts to reflect on trainings from the State, attempting to make initial sense of new information. All four cases found reports of strong outside-of-district social support systems, or capacity. Evidence suggested that there was rarely a group consensus, though, and that while everyone’s
understanding would improve, the final form of the sensemaking process did not take place until staff returned to their own district and engaged in a second level of social learning. In these latter conversations, district staff would work internally, balancing the interpretations of the multi-district group with what appeared to fit best at the local level. These conversations influenced by personal beliefs and worth-related factors.

Both Coburn (2001) and Seashore (2005) made note that strong social networks do not necessarily lead to second order types of change, and this was evident across Districts A, B, and D. The focus of the interpretation was on how to bridge parts of the TEM that were aligned with personal beliefs and that provided some sort of worth or benefit. In other words, identifying those parts of the reform that were useful locally became the goal. In the three districts mentioned above, there appeared to be little focus on the original purpose of the policy behind the TEM, and there was not much regard for implementation as prescribed by the State. District C was much more aligned with policy goals and the prescribed approach in its interpretation of the TEM; however, this did not appear to stem from social learning as much as it did from the personal beliefs of central office staff.

**Interaction Between Variables.** Each of the variables detailed above, the perceived benefit or worth of the reform to the local district, competing factors, past experiences and knowledge, personal beliefs and intent, and the quality of social interactions appeared to make unique contributions when it came to understanding how sense was made of the TEM. None of these variables by themselves, would appear adequate in describing why or how the final interpretation came about. Instead, there seemed to be a complex and uniquely local interaction between the variables that cumulatively added up to the final interpretation.
Central office staff talked and provided more information about the perception of worth- or value-related factors (importance, usefulness, benefit) than any other factor in the study; and it was the perceived worth, or lack thereof, of the TEM that ultimately drove interpretation. However, personal beliefs, in this study including the local culture and norms of the district, may have held the strongest weight as to how the TEM was understood. This factor seemed to drive what was important, what was useful, and what types of gain were meaningful (perception of value- or worth-based factors). The interaction between personal beliefs and worth-related factors had a dramatic influence on the components that were bridged and those that were buffered. Beliefs also influenced social learning. While strong across-district social capacity existed, actual sensemaking did not take place until new information was processed by the within-district social group, where reform related information was tempered by the culture and beliefs of each district. Together, these three factors, personal beliefs, worth, and social interaction, seemed to culminate and lead in all four cases to the resulting interpretation of the TEM. While other variables such as competing factors, knowledge of reform, and intent to implement were noted, none appeared to play a major role in sensemaking in any of the cases.

**Cross Case Analysis Summary.** Merriam (2009) described multiple case study designs as a method that allowed for a more compelling argument as to what was going on. Stake (1995) and Stake (2005) noted how instrumental case study design allowed for a focus on a phenomenon more so than a specific case. Yin (2003) wrote of how using a purposeful sample of cases to explore similarities in a phenomenon was advantageous. The cross case analysis in this study shows how a purposefully sampled set of four districts worked through the sensemaking process. Specifically, factors related to the perception of value or worth, personal beliefs, and social interaction could be tracked across all four cases as meaningful contributors to
sensemaking. More striking; however, was the complexity of the interactions between these three factors, and how this interaction seemed to play the greatest role in leading to how sense was ultimately made in each case.
Chapter Five: Conclusions

The purpose of this study was to explore how factors specific to the local school district interacted to influence the sense central office staff made of performance-based school reform efforts. The exploration was important because the majority of the research base on such reform has focused on technical qualities of the evaluation instruments like psychometric reliability and validity. While some have noted the role central offices play in the interpretation (sensemaking) process as a reform enters a district, little is known about how factors related to the local context influence such a process, and even less is known about how these factors might interact to explain local response. The study was based on the concept that variables related to the reform itself, specifically reliability and validity, as well as the context wherein the reform is implemented (the local district) contribute to explaining how well aligned the actual practice of reform will be to the original intent of the policy behind it.

Given the wealth of information pertaining to variables related to reform models and their various components, the current study looked to fill a gap in understanding the role of the local context. Specifically, this study explored factors related to perceptions of reform value or worth, competition for resources, past experience, knowledge, personal beliefs of central office staff, and social interaction. Four school districts were purposefully sampled in this qualitative cross-case study. Results of the study found that perception of value- or worth-related factors, personal beliefs, and social interactions contributed the most to local interpretations of the Teacher Effectiveness Model (TEM). Each of these three factors was important in all four cases studied, but the interaction between the three proved most influential. Results suggest that these co-contributing variables are vital to understanding why certain parts of the TEM were
embraced, and bridged to existing district practice, or avoided, and buffered away from district processes to the greatest extent possible.

**Discussion of Findings**

The actual process of interpretation is a cognitive constructivist activity (Spillane, 2000). In order to understand this process, the factors that influence how the individual or group at the central office level makes sense of new information must be considered. Spillane, Reiser, and Gomez (2006) frame these factors under the general themes of background knowledge and social process. Additional research suggests factors related to worth (Seashore-Louis & Dentler, 1998) and competing factors (Firestone & Shipps, 2005) might also influence interpretation. The current study supports the importance of these factors, but also indicates some meaningful caveats.

Spillane, Reiser, and Gomez (2006) use the concepts of mental schemas, assimilation and accommodation, and depth of knowledge to explain the broader theme of background knowledge. Schemas are a pre-existing mental guide used to understand new information (Spillane, Reiser, & Gomez, 2006). The concept of mental schema was found most profoundly in District C, where the prescribed approach of the TEM fit the mental structure of the Director of Curriculum and Instruction. Other explicit evidence of how mental schemas influenced interpretation was limited, but the general concept can be found by taking into consideration the personal beliefs many central office staff had about supervision and evaluation. Every administrator interviewed believed strongly in the link between teacher growth and supervision, in effect, each had a pre-existing mental guide for what supervision should look like and what it should accomplish. The evidence collected clearly showed the overwhelming influence of these pre-existing beliefs on how the TEM was interpreted, adding additional support to the
conclusions of Spilllane, Reiser, and Gomez (2006) as well as Firestone and Shipps (2005 – who explicitly wrote of the link between personal beliefs and sensemaking). When looking to understand how variables of the local context influence the interpretation of outside information, the pre-existing beliefs, or mental schemas, of those who are making sense of the new information appeared to play an exceptionally important role.

Spillane, Reiser, and Gomez’s (2006) concept of assimilation, as well as Honig and Hatch’s (2004) concept of bridging and buffering, were strongly supported by the evidence collected. In all cases, respondents used the word “leverage” and offered insight as to how they linked “new” information from the TEM to “old” ideas already existent in the district. In each case, central office staff looked to assimilate, or bridge, TEM concepts so that local initiatives would benefit. The relationship between new and old information might also be understood in terms of Crossan et al.’s (1998) description of exploitation and exploration. Central office staff exploited what they already knew and were attempting to put into practice. At the same time, they explored the new information of the TEM as it presented the possibility of strategic renewal. Each district’s balance between new and old was somewhat different, as was their perceived need for renewal; however, the overall process across all four cases was marked in its focus on leveraging.

Both Spillane (2000) and Honig (2003) wrote about first order or form and second order or function related change, and linked these changes to assimilation and bridging and buffering; with the idea that the more a district bridged or assimilated new information to pre-existing ideas, the closer implementation would be to second order, or function related, change. In the current study, there was very little evidence of any attempt at second order or function-focused change as described in these theories, but the evidence suggested an added complexity that might
extend the concepts detailed by Spillane and Honig. While administrators focused on making meaningful change related to the culture and values of their districts (hallmarks of second order or function-focused change), this change was not aligned with administrator’s perceptions of the original purpose of the policy behind the TEM. Instead, it was related to the personal beliefs and pre-existing initiatives in the district. In this way, the staff did not change the function of internal processes to align with the purpose of the TEM, but rather changed the function of internal processes to leverage their own priorities. There was very little accommodation, a creation of new processes based on new information and associated with function focused or second order change, but there was a great deal of bridging. This finding might be considered from a political perspective in that the central office apparently acted as a small coalition within the structure of Pennsylvania’s education system in order to manipulate the TEM for their own interests. Given this observation, additional investigation as to the understanding of the bridging and buffering continuum (see Honig & Hatch, 2004, p 24), local leveraging, and the role district self-interest plays in sensemaking might be in order.

The amount of knowledge and experience an individual has with reform-related concepts has been linked to form or function related change (Honig, 2003; Spillane, 2000; Spillane, Reiser, & Gomez, 2006). In the current study, this factor did not appear to be as meaningful in predicting central office response. While there was evidence that administrators had a solid understanding of some components of the TEM and had to work hard with colleagues to figure out other components, the influence of personal beliefs appeared to overwhelm any meaningful variance that depth of knowledge might contribute. The cases of Districts C and D showed that the experience of the Superintendent could influence what might be bridged and buffered, but even this appeared to be linked more to personal belief than background knowledge in and of
itself. As with above, this finding might extend our thinking on the role of background knowledge and experience, confirming that as a contributing factor to sensemaking background knowledge should be considered, but clearly there are cases, at least at the central office level, where other factors are much more meaningful.

Competing factors were noted as potential influences on sensemaking by Firestone and Shipps (2005) but not given much attention in the Spillane, Reiser, and Gomez (2006) framework. In the current study, competing factors were found in all four districts; but in none of the cases did these appear to have much of an impact on how the TEM was interpreted. The legal mandate of the TEM might have confounded the situation enough that while the reform certainly taxed local districts, actual implementation of some sort was understood as necessary. What that implementation would look like had very little to do with other factors taking time away from figuring out the TEM, and much more to do with the other factors detailed here – personal beliefs, social processes, and worth-related variables. As such, Firestone and Shipps’ (2005) suggestion of the role competing factors play in interpretation might possibly be limited to new information that is not mandated.

Spilllane, Reiser, and Gomez (2006) complete their framework of factors that influence sensemaking by considering the role of social learning. The process of taking in new information, considering it within a small group, refining, discussing, and over time transferring into permanent practice is well documented and understood (Crossan, Lane, & White, 1997; Seashore-Louis & Dentler, 1998; Spillane & Thompson, 1997; Spillane et al., 2002). The current study found this process to play out as theorized, but to also be influenced by other factors. Staff in all four districts studied commented on how poorly the TEM was rolled out by the State. As such, Assistant Superintendents and Directors of Curriculum had to meet across
districts to discuss and figure out what certain components meant. However, moving from the figuring out stage at this level to the figuring out stage back in each individual district was influenced both by social interactions in-district as well as the personal beliefs of staff and the perceived worth of each component. The purpose of the current study was not to delve deeply into social processes, and as such, there is a limit to the interpretation beyond confirming previous results pointing to the role the social process plays in sensemaking and perhaps suggesting the depth of complexity and multiple levels (across the district, within the district) in these social relationships.

Spillane, Reiser, and Gomez (2006) do not include perception of value- or worth-related factors in their frame of what influences the sensemaking process; however, an earlier work by Seashore-Louis and Dentler (1998) suggested that importance, usefulness, and gain or benefit could play a role in how central office staff respond to external, reform related information. Evidence from the current study overwhelmingly supports this idea. In all four cases, the way in which the district made decisions as to what to bridge and what to buffer was directly linked to what was important locally, what was perceived to be useful, and what might lead to some type of gain. These decisions were often linked to how a TEM component could be leveraged as detailed above. These three variables represented the way in which administrators talked about their decision making and the way in which they framed their thoughts on the TEM. Given the profound strength of perception of value- or worth-related factors in each of the four cases studied, it may prove prudent to explore how the perception of value or worth might expand Spillane, Reiser, and Gomez’s (2006) framework.

While each of the factors detailed above, particularly pre-existing schemas and personal beliefs, social processing, and perception of value or worth, were strong contributors in how
sense was made of the TEM, the interaction between these variables seemed to be the key to understanding why any given district went in the direction that it did. Evidence was often provided in terms of perception of value- or worth-related factors; however, the source of such worth was clearly influenced by the personal beliefs of the central office staff. These beliefs, in turn, were fine-tuned through social interaction in the district. Interestingly, social interaction outside of the district helped staff better understand the meaning of certain TEM components, but a deeper understanding developed in-district when that social learning was crossed with the local worth-related factors and personal beliefs.

Spillane (2000), when studying teacher implementation of math reform, used the concept of a regression equation to explain how the variance contributing to implementation might be understood. He wrote, “In this equation…each variable would be interdependent rather than independent in that the worth of any one variable in predicting implementation would depend in some part on the other variables” (p. 166). Spillane’s concept fits the results found in the current study, as well; however, the complexity and near impossibility of quantifying the variables would make an actual equation unlikely. Instead, as presented in Figure 1, a set of concentric circles might best demonstrate the relationship between influential factors and how they individually and collectively lead first to the sense that is ultimately made and then to decisions regarding whether to bridge or buffer at the local level.
Limitations

Of course, the findings of the current study must be considered in light of a number of limitations. The qualitative nature of the study makes generalization beyond the four districts of study unreasonable. While strong themes presented across all four cases, the reader must be careful not to confuse a deliberate focus on the sensemaking phenomenon with statistical ideals linked to samples and populations. All organizations have differences, and the design of the current study can only suggest that these differences be closely examined.

The factors examined in the study stemmed from a review of a rather limited research base (Honig & Hatch, 2004; Rorrer, Skrla, & Scheurich, 2008). As such, it is possible that the research design could have missed other, yet to be identified factors that influence how sense is ultimately made. While the interview process was open ended and made every attempt to explore the possibility of finding other factors, there is always the possibility that additional
sources of meaningful variance exist. The research design purposefully did not look at any district with outstanding capacity limitations, as indicated through budgeting information and school performance profiles (SPP). As a result, there is no indication as to how well the proposed model in Figure 1 might fit when capacity related factors are involved. Finally, the research design did not directly investigate how the structure of the TEM itself might have influenced how sense was made, although the idea was addressed somewhat through the investigation of worth-related factors. As is detailed in chapter two, there are a great number of concerns with the reliability and validity of the components that make up the TEM, and it is reasonable to wonder if a more competently designed model might have influenced local response differently.

One additional limitation discovered through the research was the grandiosity of all that can happen at the local level in terms of what influences interpretation and leads to how a reform is implemented. The study of school reform relative to consequence validity, or validity in use (Fuller & Hollingworth, 2014), has focused largely on what happens in the local context after a reform was implemented, from changes in curriculum, instruction and professional development (Lane, 2014) to the impact results have on different subgroups (Lane, 2014), to survival based strategies such as teaching test taking skills (instead of academic content) or even cheating on high stakes tests (Hill, 2009; Jennings & Bearak, 2014). In contrast, the current study focused on what happens in the local context before reform is implemented – the point where new information is first confronted and considered by central office staff. The distinction between before and after implementation consequences is important, as the change that takes place during the initial learning (before implementation) creates an additional variable that influences any later response (after implementation). In effect, how sense is made at the central office when a
reform is first introduced would create another level of variables necessary for consideration when trying to understand the full influence on how reform-related models link back to policy intent. In the current study, the before implementation process is investigated, with results suggesting the importance of personal beliefs, perception of worth, and social interaction. The study does not go on to explore how each of these factors, or their cumulative interaction, influence response across the district during or after implementation of the TEM, but it is reasonable to wonder how the former might impact the latter. Clearly, additional research as to how the sensemaking process might play out with building level principals and classroom teachers could complement the current research and hopefully provide additional insight as to the multiple and complex sensemaking forces at work.

**Future Directions**

There are some marked results from the current study; each of which has implications for stakeholders and sets the stage for additional investigation and research. First, perception of value- or worth-related factors and personal beliefs emerged as factors from the local context that play heavily into sensemaking. These factors have been discovered in past research, but do not seem to have been given as much attention as background knowledge and social interaction. Findings of this study suggest perception of value or worth and personal beliefs warrant more attention. Second, the interactions between all local factors, but especially the perception of value or worth, personal beliefs, and social interaction were profound. While very complex, and thus, difficult to isolate and study, the interaction phenomenon appears very important to how sense is made. This should be recognized and investigated further.

The perception of the value- or worth-related factors of importance, usefulness, and benefits identified by Seashore-Louis & Dentler (1998) were overwhelmingly important to how
central office staff made decisions regarding what to bridge and what to buffer when it came to the teacher effectiveness model (TEM). While perception of value- or worth-related factors have not been given a great deal of attention in more recent works on sensemaking (Spillane, Reiser, & Gomez, 2006), the current findings would seem to justify an exploration into expanding the sensemaking frame beyond just background knowledge and social interaction. Future research might stem from these findings. Seashore-Louise and Dentler (1998) provided a comprehensive design and investigation into value- or worth-related factors. The current study confirms findings that these factors influence how sense is made of outside information. Additional study is warranted. Future research may look simply to replicate findings in order to add weight to the idea that the perception of value or worth deserves to be included in any sensemaking theory just as predominately as background knowledge and social interaction already do.

Factors related to personal beliefs (Firestone & Shipps, 2005) played a vital role in how sense was made across the four cases studied. Personal beliefs can be framed as pre-existing mental understandings, or “schemas,” one of the three notable factors within background knowledge as explained by Spillane, Reiser, & Gomez (2006); however, the degree to which personal beliefs influenced the perception of value- or worth-related factors suggests that this category warrants additional attention and potentially special recognition of its role in the sensemaking process. The current study gave only limited attention to personal beliefs when initially designed. The qualitative process of data collection allowed for additional probing into the topic as it emerged; however, additional research might specifically plan to address personal beliefs, looking at beliefs as they relate to both to specific reform but also to education in general, and other topics that might influence reform – such as local control and accountability.
Results related to the perception of value- or worth-factors as well as personal beliefs should be noted by policy makers and those who shape reform design at Federal or State levels. As is well documented (Honig & Hatch, 2004), local districts make decisions as to what information to keep (bridge) and what information to minimize (buffer). Current results suggest that these decisions are heavily influenced by what the local district considers to be of value or worth. This, in turn, is influenced by personal beliefs. In order for Federal and State policy to add meaning to the local level, as opposed to simply becoming an additional bureaucratic necessity (form-focused or first order change – Honig, 2003; Spillane, 2000), policy makers and those who shape reform models might reconsider recent one-size-fits-all types of reform; knowing that what is valuable to one district will not be in another. Policy with greater flexibility and that which takes local differences into consideration might evidence understanding for the factors that influence the bridging and buffering phenomenon.

The interaction between factors, especially the perception of value or worth, personal beliefs, and social interaction - was marked. Such interaction has been observed (Spillane, 2000) but limited attention has been given to how factors actually interact and influence each other, as well as the end result of sensemaking. Considering future research, as noted above, the interaction phenomenon is complex, and a prescribed quantitative method (such as multiple regression) would be difficult; however, the interplay found in the current study would seem to suggest that additional exploration into factor-related interaction is worthwhile. Because the study of how variables interact is dependent on the variables, much like the current study, any research on interaction would likely need to first figure out what was happening with each factor and then specifically focus on how these interacted with each other.
Policy makers and those who shape reform-related models might take note of the complexity introduced by the concept of interaction. Evidence from the current study finds that it is not just the perception of value or worth a reform might bring (or not bring), the personal beliefs of central office staff, or the quality of social interactions, but rather it is each of these as well as how they interact and influence each other that leads to the ultimate sense that is made of reform related information. With this in mind, overly prescriptive policy or models that are intended to “fix” large scale issues in education, such as increasing student achievement by evaluating teacher effectiveness, might be understood at best as somewhat quixotic, and at worst, as at risk to lead to great resource expense for what ultimately will provide very little value added.

**Summary**

The current study set out to explore the local factors that influence how a reform is interpreted within a specific school district. Findings support the role of the perception of value or worth, personal beliefs, and social interactions. These findings are consistent with prior knowledge, but add to the field an emphasis on perception of value or worth and personal beliefs, two variables that have received limited attention. Findings also add to the field the importance of considering the interaction of factors when studying sensemaking. Opportunities for future research are abundant and justification for federal and/or state policies that are more sensitive to local differences is provided.
References


added models. [Policy Information Perspective]. New Jersey: ETS.


Kane, T. J. & Staiger, D. O. (2002). The promise and pitfalls of using imprecise school
accountability measures. *Journal of Economic Perspectives*, 16(4) 91-114.


http://www.pdesas.org/Instruction/StudentLearningObjectives/

Pennsylvania Department of Education (n.d.). *Overview of Pennsylvania’s Approved No Child Left Behind Waiver*. Retrieved from

http://www.portal.state.pa.us/portal/server.pt/community/federal_programs/7374/p/1433522

Pennsylvania Department of Education. (2015). School Performance Profile data, retrieved from

http://paschoolperformance.org/.


http://www.portal.state.pa.us/portal/server.pt/community/educator_effectiveness_project/20903


http://www.portal.state.pa.us/portal/server.pt/community/pennsylvania_department_of_education/7237/p/1604316


http://www.portal.state.pa.us/portal/server.pt/community/pennsylvania_department_of_education/7237/p/1604316


http://www.portal.state.pa.us/portal/server.pt/community/pennsylvania_department_of_education/7237/p/1604316

Pennsylvania House Bill No. 1901 of the Session of 2012.


measures into teacher evaluation systems. RAND.


http://www.portal.state.pa.us/portal/server.pt/community/federal_programs/7374/p/14335

22


EVAAS-Statistical-Models.pdf


CA: Sage.

Sage.

Sage.
Appendix A: Interview Protocol

Interview Questions

Introduction
- Thank you for your time today. I would like to ask you about the process of bringing the Teacher Effectiveness Model into this district. As we have already discussed, I will be recording the interview so as that I do not miss anything you tell me. You are not obligated to answer anything you do not want to and you may end the interview at any time. Do you have any questions before we begin?

Research Questions¹
- From your perspective, tell me how teacher evaluation has changed in the past few years.
- What would you say are the three biggest issues related to the new teacher effectiveness model for your district right now?
- What do you believe are the strengths of the new teacher effectiveness model? What do you believe are the weaknesses?

Closing
- Is there anything else about the Teacher Effectiveness Model that you would like to add?
- Is there anyone else in the district who was involved with the process and would you mind if I talked with them?
- Thank you for your time today. As I begin to review this information I may ask you to look over the notes I’ve taken to make sure I’ve captured what you intended to tell me; is that okay?

¹ Each Research Question may be followed up by probing questions including requests for examples, elaboration, further explanation, or clarification.
### Appendix B: Framework for Data Collection

<table>
<thead>
<tr>
<th>Research Question</th>
<th>Propositions</th>
<th>Level 2 Questions</th>
<th>Possible Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>How have central offices made sense of the Teacher Effectiveness Model?</td>
<td>Success (Seashore-Louis &amp; Denter, 1998)</td>
<td>• What is the general status of the context/district of the central office?</td>
<td>• Archival records including district webpage, the SPP, the NCES, budget documents</td>
</tr>
<tr>
<td></td>
<td>Degree to which reform is important, useful, or will lead to a gain (Seashore-Louis &amp; Denter, 1998).</td>
<td>• Was the EEM important at the local context?</td>
<td>• Interviews</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Did the EEM seem to help solve any local problems or concerns?</td>
<td>• Documents (?)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Was the EEM perceived to lead to anything beneficial for the district or individuals within?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Competing factors (Firestone &amp; Shipp, 2005) &amp; limited resources (Ogawa et al., 2003; Olsen &amp; Sexton, 2009).</td>
<td>• What are the pressing political factors in the district? Is there evidence that they are incompatible with the EEMs?</td>
<td>• Archival records including media reports, school board agendas/minutes, district or other local organization, webpages, budget documents</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• What are the pressing bureaucratic (system) factors in the district? Is there evidence that they are incompatible with the EEMs?</td>
<td>• Interviews</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• What are the pressing professional factors (union, other) in the district? Is there evidence that they are incompatible with the EEMs?</td>
<td>• Documents including meeting agendas, memos</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Are there other pressing factors in the district that might compete or be incompatible with the EEMs?</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• What evidence is there to judge the limitations of physical capacity or other resources?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Past experience (Spillane, 2000; Spillane &amp; Thompson, 1997) and knowledge of reforms (Coburn, 2005).</td>
<td>• What is the perceived history of reform in the district?</td>
<td>• Interviews</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• How well do staff in the central office understand the EEMs?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Personal beliefs (Firestone &amp; Shipp, 2005) and intent to implement (Sanders, 2012)</td>
<td>• What are the personal beliefs of central office staff related to the EEMs?</td>
<td>• Interviews</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Was there knowledge of and an intent to implement the EEMs in a way that would identify and remove ineffective educators?</td>
<td>• Documents including meeting agendas/memos</td>
</tr>
<tr>
<td></td>
<td>Quality of social interactions (Coburn, 2001)</td>
<td>• Who was a part of the social network trying to understand the EEMs?</td>
<td>• Interviews</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• What was the perception of the input from others inside or outside the district relative to the EEMs?</td>
<td></td>
</tr>
</tbody>
</table>
VITA

Andrew E. McCrea
70 McCrea Rd
Newville PA 17241

Education
- The Pennsylvania State University, State College PA
  - Ph.D. in Educational Leadership, December 2015
  - Superintendent’s Letter of Eligibility, May 2014
- Shippensburg University, Shippensburg PA
  - Supervisory Certification in Pupil Services, December 2009
- Millersville University
  - Certification in School Psychology, June 2004
  - M.S. in Psychology, August 2003
  - B.A. in Psychology, May 2001

Professional Certifications
- Pennsylvania Certification Superintendent’s Letter of Eligibility
- Pennsylvania Certification Supervisor of Pupil Personnel Services PK-12
- Pennsylvania Certification School Psychologist PK-12
- National Certification - School Psychologist

Work Experience
- Director of Pupil Services, Lower Dauphin School District (June 2011 – Present)
- Adjunct Professor, Wilson College (January 2010 – May 2010)
- Research Assistant, The Pennsylvania State University Medical Center & College of Medicine (August 2002 – June 2004)

Publications & Written Works

Presentations