THE DIFFERENCE IN DISTRIBUTED PRINCIPAL LEADERSHIP PRACTICES ACCORDING TO THE COMPREHENSIVE ASSESSMENT OF LEADERSHIP FOR LEARNING BETWEEN TEACHERS AND SCHOOL LEADERS, GRADE LEVEL AND REGION

by

Margaret A. Freeman

Dissertation Committee:

Professor Alex J. Bowers, Sponsor
Professor Jeffrey Young

Approved by the Committee on the Degree of Doctor of Education

Date May 20, 2020

Submitted in partial fulfillment of the requirements
for the Degree of Doctor of Education in
Teachers College, Columbia University

2020

ABSTRACT

THE DIFFERENCE IN DISTRIBUTED PRINCIPAL LEADERSHIP PRACTICES ACCORDING TO THE COMPREHENSIVE ASSESSMENT OF LEADERSHIP FOR LEARNING BETWEEN TEACHERS AND SCHOOL LEADERS, GRADE LEVEL AND REGION

Margaret A. Freeman

Leadership for learning has arisen as a robust framework for measuring principal leadership practices that impact student achievement by synthesizing fundamental theories of principal leadership, such as instructional leadership, transformational leadership, and shared instructional leadership. Research affirms the need for principals to prioritize leadership for learning practices based on the context. However, we know little about how context impacts which leadership for learning practices are prioritized by teachers and administrators in these different contexts of school level (elementary or secondary) and region (rural, small town, suburban, urban). The purpose of this dissertation study is to examine how differences in context affect specific leadership for learning practices. These differences will provide evidence for the need to revamp the principal evaluation systems that are currently in use and allow us to identify opportunities for preparation, support, and professional development for principals. I use the Comprehensive Assessment of Leadership for Learning (CALL) survey for this dissertation study as it combines recent research on principal leadership to measure

specific leadership skills. Collected by the researchers of the CALL group at the University of Wisconsin-Madison, I use the 2016-2019 CALL dataset in the present study, which includes 281 schools with 536 administrators and 8,027 teachers. This dissertation study examines a) differences in teacher and administrator ratings within the CALL assessment, b) differences in secondary and elementary school level ratings within the CALL assessment, c) differences in administrator's ratings by region within the CALL assessment. The results of this dissertation study show that a) administrator perceptions of the frequency of leadership for learning practices are always higher than teacher scores for any leadership for learning practice, b) elementary schools had fewer disagreements across the leadership for learning practices in comparison to secondary schools when comparing the perceptions of the frequency of practices in leadership for learning scores between teachers and administrators, and c) substantial differences occur in the perception of the frequency of leadership for learning practices between urban school administrators and administrators in other regional contexts (rural, small town, suburban), especially around issues of school climate, summative evaluation of teachers and students, and resourcing strategically.

©Copyright Margaret A. Freeman 2020

All Rights Reserved

DEDICATION

I dedicate this dissertation to my 457 + 2 children. First, I dedicate this to the children of Jersey City who have changed my life by exposing me to a world that is so different than the one that I experienced growing up in a suburban community. Your stories and your perseverance push me to be a better leader. Second, I dedicate this to my own two children, Adele and Calvin, who fill me with joy and hope and are the inspiration behind the work that I do every day as a principal.

ACKNOWLEDGMENTS

I would like to thank the CALL team for their generosity and support throughout this dissertation process. First, thank you for creating such a practical instrument that can be used by principals across this country to improve instructional outcomes for students. Second, I sincerely thank Dr. Richard Halverson and Dr. Mark Blitz for allowing me to use their data for this dissertation. Dr. Blitz spent several hours on the phone with me in the past two years, explaining how to read and use this data. Thank you, Dr. Halverson and Dr. Blitz, for allowing me to complete this important research using data that you collected.

Additionally, I would like to thank Alex J. Bowers, Jeffrey Young, Sonya Douglass Horsford, and Kevin Dougherty for agreeing to serve on my oral defense committee. I am especially grateful to Dr. Bowers who taught me to understand statistics, which is something I never thought possible. Thank you for the countless hours you spent over the past three years to help shape my ideas, improve my writing, and guide me through this process. Thank you for teaching me how setting high expectations can inspire students to greatness. Thank you to Dr. Young, who taught me about the importance of learning your context. Thank you to Dr. Horsford, who guided my research ideas very early on in this process and then agreed to be on my committee at the end.

Thank you to the One who orders my steps and holds me in the palm of His hand. This journey was only possible because of Your grace. I owe a debt of gratitude to my mother, Lois, and father, Dave, for watching my children, for proof-reading my drafts, and for always believing in me. I am grateful to my husband, Joel, for his support throughout this process and for always encouraging me to be my best self. MAF

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter I

INTRODUCTION

Problem Statement

By synthesizing key theories of principal leadership, such as instructional leadership, transformational leadership, and shared instructional leadership, leadership for learning has arisen as a robust framework for measuring principal leadership practices that impact student achievement. In his study of the leadership for learning framework, Hallinger (2011) writes:

> Leaders who possess a single set of tools will find themselves bouncing around from success to failure without understanding why. The capacity to read your context correctly and adapt your leadership to the needs largely determines your success…There is no one best leadership style for fostering learning in schools. We are learning more and more about the ways that leaders need to match strategies to contexts; more research on this point is needed. (p. 14)

This study is in answer to the need posed as it examines how leaders match specific leadership for learning practices in different contexts. The purpose of this dissertation study is to examine how differences in context, as defined as school level and region, affect ratings on leadership for learning practices. These differences will provide evidence for the need to revamp the principal evaluation systems that are currently in use as well as allow me to identify opportunities for preparation, support, and professional development for principals.

It is essential to measure principal leadership practices due to the impact that principals have on student achievement and school climate (Bowers & White, 2014; Brown & Wynn, 2009; Day, Gu, & Sammons, 2016; Greenlee & Brown Jr., 2009;

Johnson, Kraft, & Papay, 2012; Leithwood, Louis, Anderson, & Wahlstrom, 2004). There is a need for measures of principal leadership that will examine principal leadership practices more specifically in terms of their context, as we know little about how context affects how principals prioritize different leadership tasks (Hallinger, 2011; Halverson & Kelley, 2017). This dissertation study will use the Comprehensive Assessment of Leadership for Learning (CALL) to explore how different contexts affect principal leadership practices by comparing the mean scores of teachers and principals (Day et al., 2016; Hallinger, 2005; Halverson & Kelley, 2017; Southworth, 2002; Urick & Bowers, 2014b). I define context for the present study as the school level and region. The CALL measures the distributed leadership of school administrators and teachers on five research-based domains that are divided into twenty subdomains (Halverson & Kelley, 2017; Halverson, Kelley, & Shaw, 2014). I will be able to examine how teachers and administrators prioritize leadership differently by comparing domain ratings from their responses on the CALL survey. By separating the data by school level and region, I will examine how context affects the prioritization of those leadership tasks. Through identifying where these differences occur, I will also highlight key findings for administrator preparation, support, and professional development.

Three main research concepts guide this dissertation study. First, recent educational research literature has shifted from traditional theories of leadership to integrated theories of leadership that highlight key actions and practices of principal leaders and how they affect student and teacher learning (Cosner & Jones, 2016; Day et al., 2016; Hallinger, 2011; Leithwood & Jantzi, 2008; Leithwood, Leonard, & Sharratt, 1998; Nguni, Sleegers, & Denessen, 2006; Spillane, Halverson, & Diamond, 2004). The

three main traditional theories that have dominated the research around principal leadership are instructional leadership, transformational leadership, and shared or integrated instructional leadership. The specific leadership practices that have had an effect on student achievement across the three theories (instructional, transformational, shared) has led to the development of a leadership for learning framework (Boyce & Bowers, 2018b). Based on forty years of research, this framework highlights five key dimensions of leadership for student improvement (Hallinger, 2011). The leadership for learning framework provides the research base for the CALL survey, which includes survey questions that refer specifically to research-based effective school leadership practices (Halverson & Kelly, 2017).

Second, while the focus of this present study is a comparison of administrators' scores based on two specific contexts, I discuss current methods of principal evaluation. As Fenton writes, "the goal of principal evaluations is to improve principals' instructional leadership capacity" (2016, p. 20). However, current principal evaluation models fail to acknowledge and evaluate the role that context plays in principal leadership, limiting their ability to increase principal leadership capacity (Agasisti, Bowers, & Soncin, 2019; Day et al., 2016; Halverson & Kelley, 2017; Leithwood & Riehl, 2005; Leithwood & Sun, 2012; V. M. Robinson, Lloyd, & Rowe, 2008). Many current models also use quantitative measures of principal leadership, such as Value-added Measures (VAM's), that are statistically inaccurate in measuring student and school academic growth (Bowers & White, 2014; Close et al., 2018; Fuller, Hollingworth, & Liu, 2015; Piro, Wiemers, & Shutt, 2011). A key reason for these inaccuracies is those principal evaluation methods are driven by general federal and state policies that use general

rubrics, which do not align with instructional leadership activities that consider the specific context (Hallinger, Heck, & Murphy, 2014). This dissertation study is a descriptive quantitative study. I discuss principal evaluation methods because the newer principal evaluation methods stand in contrast to previous evaluation systems. The findings of the current study will provide evidence for this contrast and an opportunity to examine how current practices can be changed to provide principal leadership growth.

Third, I discuss the research base used for the CALL survey, so there is an understanding of the data collection method used in this dissertation study. The CALL combines recent research on principal leadership and effective principal evaluation methods in a formative assessment approach designed to improve teaching and learning through measuring leadership for learning practices (Mark H. Blitz & Modeste, 2015; Halverson et al., 2014). Written through a grant funded by the Institute of Educational Sciences in the US Department of Education, the CALL survey uses the leadership for learning framework to measure leadership practices in five domains. Principals take the survey to assess their competency in the following five domains: *Domain 1: Focus on Learning*, *Domain 2: Monitoring Teaching and Learning*, *Domain 3: Building Professional Community*, *Domain 4: Acquiring and Allocating Resources*, and *Domain 5: Establishing a Safe and Effective Learning Environment*. Teachers and staff also have the opportunity to complete the survey, measuring the leadership of the building administrators in the same domains (Blitz & Modeste, 2015; Halverson & Kelly, 2017). By comparing scores of teachers and administrators, schools have a complete picture of how leadership is exercised and perceived at the school, given the school's unique context (Halverson & Kelly, 2017). An administrator may believe that they understand a

particular practice, but their perspectives may or may not align with their teachers. If the purpose of leadership is to make an impact on its followers, how leaders and teachers perceive the same domains or skills is essential to examine the effectiveness of leadership in those particular areas (Bowers, Blitz, Modeste, Salisbury, & Halverson, 2017; Boyce & Bowers, 2018a; Reich, 1988; Urick & Bowers, 2014b, 2019)

Significance of the Study

There is significant research on the indirect effects of principal leadership that shows their actions influence school and classroom conditions and teacher-perceived student outcomes (Boyce & Bowers, 2018b). This dissertation study will add to that research by examining the principal's perceptions of their leadership (Hallinger, 2003, 2005; Hallinger & Heck, 1996; Leithwood, 1994; Leithwood & Jantzi, 2005; Southworth, 2002; Urick & Bowers, 2011, 2014a). As noted by Blitz and Modeste in their 2015 paper, "there is value in evaluating the individual principal in order to get a sense of that individual's performance. There is also value in gauging the effectiveness of leadership by focusing on the work of school leaders across a given school." The present study extends Blitz and Modeste's research by using similar methods to evaluate where teachers and administrators may differ in their ratings of the effectiveness of leadership for learning practices. In this dissertation study, administrator perceptions of the frequency of leadership for learning practices are always higher than the scores of teachers for any given subdomain. The differences between these groups aids in understanding how leadership is perceived to show that administrators and teachers do not always agree on the frequency of the leadership practice (Bowers et al., 2017; Urick

& Bowers, 2019). I also demonstrate that there are significant differences due to context, particularly by the school level and between urban administrators and administrators in other regional contexts.

This study makes contributions to both theory and practice. This research adds to theory by:

a) illustrating the impact of the context of school level on the amount and type of instructional leadership the principal delivers differs based on the size of the school (Hallinger, 2018)

b) examining the context of region to find that 5% of the variance is accounted for by the difference by region, specifically between the urban administrators and the administrators of the other regional contexts (rural, small town, suburban).

As Halinger (2018) notes, "scholars have conducted empirical studies of leadership in urban and rural schools for close to 100 years. Yet, there have been no systematic reviews that carefully analyze and distinguish the types of leadership strategies that characterize these different contexts." This dissertation study provides evidence that the leadership strategies of urban administrators are different than their counterparts in other communities. Urban administrators rated themselves differently in six leadership for learning subdomains around the areas of school climate, resourcing strategically, and summative evaluation. There is much qualitative research on urban leadership requiring different skills. Still, this dissertation study provides significant quantitative results that urban principals prioritize leadership differently around student discipline, resources, and summative evaluation (Khalifa, Gooden, & Davis, 2016; Theoharis, 2007).

The results of the present study also provide useful implications for practice and policy. First, administrator means are always higher than teacher means for any given subdomain, which illustrates the need for administrator evaluation systems that include input from teachers, such as those that incorporate a 360-degree framework (Bowers et al., 2017; Goldring, Huff, May, & Camburn, 2008; Liu, Springer, Stuit, Lindsay, & Wan, 2015; Shoaf, Zigler, & Beebe, 2013; Tran & Bon, 2015). Second, previous literature has shown that the data used for these methods of Value-added Models is statistically biased, particularly against low performing schools (Close et al., 2018; Jason A. Grissom & Susanna Loeb, 2011; Orr, 2013). The data from this dissertation study suggests that the rubric component of principal evaluation systems may also create bias as urban principals rate themselves distinctly different from the other three groups in research-based leadership practices. Principal evaluation systems might benefit from a goals-based system that is unique to each location (Fuller et al., 2015; White, Makkonen, Vince, & Bailey, 2012; Williams, 2015). Third, since urban administrators rated themselves different on six specific leadership tasks, administrator preparation programs might seek to use this research to provide targeted support to those administrators planning on working in an urban context. Fourth, while there was a significant difference in means in both comparisons of administrators (school level, region), no one group had consistently higher mean scores. The variance in higher mean scores indicates that principals from different contexts are not necessarily better trained or prepared; they just prioritize different leadership tasks. State departments might use the data to create professional learning communities where principals would be able to share strengths and weaknesses in a critical friends network with the purpose of growth and learning from each other.

Research Questions

These contributions and the above issues are addressed through the following research questions:

1. To what extent do teacher and administrator perspectives differ based on ratings within the CALL assessment?
2. To what extent do elementary and secondary school administrator perspectives differ based on ratings within the CALL assessment? To what extent do elementary teacher and administrator perspectives differ based on ratings within the CALL assessment? To what extent do secondary teacher and administrator perspectives differ based on ratings within the CALL assessment?
3. To what extent do administrator's ratings differ between rural, small town, suburban, and urban schools based on ratings within the CALL assessment?

Overview of Methods

This dissertation study is a secondary analysis of CALL data. Different districts and states have chosen to contract with the CALL to take this survey. As part of their participation in the CALL, they agree that the de-identified data may be used for future research. The researchers at the University of Wisconsin-Madison collected this data and shared it with me. I use the CALL for this dissertation study due to its unique framework, which measures specific leadership skills in a distributed fashion (Halverson et al., 2014). The individual multiple-choice items highlight the degree to which teachers and administrators are both involved in each of the tasks through their structured responses. The unique structure of these multiple-choice items measures the degree to which

administrators distribute the leadership of the task (Halverson & Kelley, 2017; Leithwood, 1994; Spillane, 2006; Spillane et al., 2004).

I first analyzed and transformed the CALL data. Participants took this survey online between 2016 and 2019. As such, I refer to the data used in this dissertation study as the 2016-2019 CALL dataset. In completing this survey, participants classified themselves as administrators or teachers. For this analysis, there are 281 schools with 536 administrators and 8,027 teachers. I averaged question scores for each participant to generate subdomain means (Blitz & Modeste, 2015; Bowers et al., 2017). I then created z-scores to standardized responses for both groups of respondents. I use z-scores to show the score's relative position or where a specific score is in relation to the rest of the scores in the distribution (Mertler & Reinhart, 2017). The z-scores allow me to compare subdomain means.

To determine to what extent distributed-leadership domains differ between teacher and school leader perspectives based on ratings within the CALL assessment, I ran twenty independent samples t-tests on administrators' and teachers' responses, one for each subdomain mean. I ran t-tests at $p = < .05$ with a Bonferonni adjustment of $p = <0.0025$ (Cohen, Cohen, West, & Aiken, 2003). I calculated effect sizes using Cohen's d (Cohen, 1992).

For the second analysis, I used t-tests to examine to what extent elementary and secondary responses differ. I separated administrator responses from the entire sample. Since respondents must identify whether or not they are elementary or secondary, I divided administrator scores into two groups based on that demographic question. I ran an independent samples t-test on each of the twenty subdomains at $p = < .05$ with a

Bonferonni adjustment of p = <0.0025 (Cohen et al., 2003). I then returned to the entire data set and split the data into two groups, secondary and elementary. I ran twenty independent samples t-tests at p = < .05 with a Bonferonni adjustment of p = <0.0025 on the split file to compare differences in means between the elementary teacher and administrator scores to and then the secondary teacher and administrator scores (Cohen et al., 2003). I used Cohen's d to calculate effect sizes for each analysis. The effect sizes allowed me to compare between the two groups to examine which group had more significant differences between means.

Lastly, to examine the extent to which principal ratings on CALL subdomains differ due to region, I ran a one-way ANOVA test at p = < .05 to examine group mean differences between subdomains. I classified administrators by their self-identified region of rural, small town, suburban, or urban. The ANOVA calculated effect sizes through partial eta squared. Next, I conducted Bonferonni's post hoc analysis to determine the differences between groups. I used Cohen's d to calculate effect sizes between the groups that had significant mean differences.

Summary of Results

The results of the present study illustrate that when measured through the CALL survey, administrators prioritized different leadership tasks in different contexts, particularly school level (elementary or secondary) and region. Overall there were five key findings. First, administrator perceptions of the frequency of leadership for learning practices are always higher than teacher scores for any given subdomain in the 2016-2019 CALL dataset. This finding replicates and extends previous research with the

CALL validation data set from 2012 (Blitz & Modeste, 2015; Bowers et al., 2017). In the present study, the largest differences indicate that principals, in comparison to teachers, more often report a focus on the three areas of *maintaining a school-wide focus on learning* (1.1), *formal leaders are recognized as instructional leaders* (1.2), and *formative evaluation of teaching* (2.3). Second, the three comparisons between administrators and teachers (whole sample, elementary, secondary) had more statistical significance across subdomains than either comparison between administrators (school level, region). Third, while this dissertation study replicates many of the findings from Blitz and Modeste's (2015) analysis of the CALL 2012 validation dataset, my analysis of the larger and more recent 2016-2019 CALL dataset, indicates that in comparing the leadership for learning domains that are most different between teachers and administrators (high disagreement), only half of the CALL subdomain teacher-administrator mean differences have a similar rank order to Blitz and Modeste (2015). The difference may be due to merely different datasets (validation versus the 2016-2019 CALL dataset), changing perceptions over time, or the different samples. Fourth, in comparing the perceptions of the frequency of practices in leadership for learning between teachers and administrators, elementary schools in the 2016-2019 CALL sample had much fewer disagreements across the subdomains in comparison to secondary schools, indicating perhaps a stronger alignment between teachers and administrators in the use of leadership for learning practices at the elementary level. Lastly, I demonstrate that there are considerable differences in the perception of leadership for learning practices between urban school administrators and administrators in other regional

contexts, especially around serving the needs of struggling students, assessment, evaluation, external expertise, student discipline, and a clean environment.

Positionality

It is important to acknowledge my own bias and positionality as a researcher. I am currently a Middle School principal in a large urban district in New Jersey. I have spent my entire career teaching and leading in a large urban school district. I have always worked at the secondary level in a middle school or high school. I feel passionate about the achievement and opportunity gaps that appear between urban and suburban schools (Khalifa et al., 2016). This is particularly the case because I grew up and currently live in a suburban environment. With a child currently in a suburban school system, I frequently note the difference between the issues that I face as a secondary principal in an urban environment and the problems that my daughter's principal faces as an elementary principal in a suburban setting.

This dissertation study is also influenced by my interest and work around teacher and principal evaluation systems. To meet the requirements under the No Child Left Behind Act, New Jersey enacted the TeachNJ act, which required a new evaluation system for teachers and principals that would incorporate Student Growth Objectives (SGO's) and Student Growth Percentiles (SGP's) through students' standardized test scores. As part of the implementation of that law, an advisory committee, Achieve NJ Advisory Committee (ANJAC), was formed. I served on ANJAC for two years and worked with the Office of Evaluation at the New Jersey Department of Evaluation on several other projects related to the implementation of the law. The experience exposed

me to the flaws, successes, and challenges of the implementation of statewide evaluation systems.

Lastly, I am currently enrolled in a doctorate program on Urban Educational Leadership. This program specifically addresses issues of social justice and areas that make urban leadership distinct. While this bias does not directly influence the methods of this dissertation study, as I will be using quantitative measures, I acknowledge that my positionality may influence perceived implications.

Definition of Terms

Effective Principal Leadership- In this dissertation study, "the ruler for judging the effectiveness of educational leadership is its impact on the learning and achievement of students for whom the leader is responsible" (Robinson, 2011, p. 4)

Instructional Leadership- leadership focused on instruction defined through three dimensions: identifying the school's mission, managing the instructional program, and promoting a positive school culture

Transformational Leadership- charismatic leadership where the purposes of the leader and the followers become fused so that they are working together towards a common cause and the leader moves their followers to a place of higher morality

Shared or Integrated Instructional Leadership- leadership defined as the active collaboration of the principal and their followers, often teachers, on curriculum, instruction, and assessment that occurs informally and formally with specific roles (Marks & Printy, 2003; Printy, Marks, & Bowers, 2009)

Leadership for Learning- a theory of leadership that combines aspects of instructional leadership, transformational leadership, and shared instructional leadership (Hallinger, 2011)

Leadership Typology – the theory that leaders follow more of a type or style, selecting particular leadership skills to meet the challenges of their context as opposed to a specific style of leadership. "Leaders enact multi-dimensional styles of leadership, in which high transformational leadership is a necessary but insufficient condition for high shared instructional leadership, and as such school personnel enact a continuum of practices across these multiple dimensions" (Bowers et al., 2017).

Comprehensive Assessment of Leadership for Learning - a formative assessment that draws on twenty years of research in school effectiveness to formatively assess principal leadership tasks across five domains, divided into twenty subdomains. While measuring these five domains, CALL incorporates two key dimensions of leadership- that leadership is as follows:

> **Socially distributed-** the principal does not lead alone. Other key actors coordinate the work in schools. The tension between leaders and followers is shown in whether these tasks are developed together or imposed upon the followers (Halverson & Kelly, 2017)
>
> **Situationally distributed-** the school has a specific context which influences how a leader prioritizes tasks

Context- how the school is situated: sociocultural context (Spillane et al., 2004), individual vs. organizational, external vs. internal, socioeconomic (Day et al., 2016; Gu & Johansson, 2013); for this dissertation study, context refers only to school level (elementary or secondary) and region (rural, small town, suburban, urban)

Region- the CALL assessment has survey participants identify their region as rural, small town, suburban, urban. For this dissertation study, region is used as the moniker for when those locations are talked about together.

Elementary or Secondary-the CALL assessment divides participants into two survey groups: elementary or middle and high school. For purposes of clarity, the middle school and high school group will be referred to as secondary.

Evaluation- the summative judgment of a principal's performance determined by the cumulative results of data, observations, and interactions (Derrington, 2018)

Chapter II

REVIEW OF THE LITERATURE

The purpose of this dissertation study is to examine to what extent leadership practices are contingent on context. Though principal leadership is multifaceted and evolving in nature, it is important to measure due to the impact that principals have on student achievement and school climate (Brown & Wynn, 2009; Day et al., 2016; Greenlee & Brown Jr., 2009; Johnson et al., 2012; Leithwood et al., 2004). While there is an emphasis on the measurement of principal leadership through principal evaluation systems, many current evaluation systems are general, failing to incorporate research on principal leadership and failing to provide specific feedback around how context affects principal leadership (Close et al., 2018; Ehlert et al., 2016; Halverson & Kelley, 2017; Orr, 2013; Sun & Youngs, 2009). Thus, there is a need for evaluations of principal leadership that will measure principal leadership practices more specifically in terms of their context. This dissertation study will use the Comprehensive Assessment of Leadership for Learning (CALL) to examine how context affects principal leadership (Halverson & Kelly, 2017). The CALL combines recent research on principal leadership in a unique framework, which measures specific leadership skills in a distributed fashion (Hallinger, 2018; Halverson et al., 2014). By using the CALL, I will be able to examine how the uniqueness of each school context might dictate how principals exercise leadership (Day et al., 2016; Hallinger, 2005; Halverson & Kelley, 2017; Southworth, 2002; Urick & Bowers, 2014b).

In this review of the literature, I will provide a connection between the key concepts behind this dissertation study:

(a) principal leadership,

(b) the effectiveness of current principal evaluation systems, and

(c) the Comprehensive Assessment of Leadership for Learning, an assessment tool for principal leadership.

Next, I will give an overview of the conceptual framework that outlines how the research regarding principal leadership has influenced the basis for my study. This section will conclude with research questions.

Review of the Literature

The role of the principal has an essential impact on the instruction and climate of the school (Bryk, Sebring, Allensworth, Easton, & Luppescu, 2010; Leithwood & Riehl, 2005; Sebastian, Huang, & Allensworth, 2017). Some research shows that aside from teachers, principals may be one of the greatest influences on student achievement (Leithwood et al., 2004). In looking at schools that made remarkable gains in student achievement, qualitative case studies on these "outliers" point to the educational leader having significant effects on student learning (Leithwood & Riel, 2005). Though case studies acknowledge this generally, quasi-experimental quantitative studies have attempted to use a number to measure this link and show that leadership may account for 3-5 percent of the variance in student achievement (Leithwood & Riel, 2005). Other research indicates that principal actions not only affect student achievement but the school climate and teacher retention (Blase, 1987; Brown & Wynn, 2009; Greenlee &

Brown Jr., 2009; Jason A Grissom & Susanna Loeb, 2011; Johnson et al., 2012; Urick & Bowers, 2014b). A principal's ability to affect school climate is higher in urban schools where principal support plays a critical role in teachers' decisions to stay, specifically in high needs schools (Bowers & White, 2014; Cosner & Jones, 2016; Greenlee & Brown Jr., 2009; Hughes, Matt, & O`Reilly, 2015; R. Ingersoll & May, 2016; Johnson et al., 2012). Thus, principals have a significant effect on schools and their student achievement (Bryk et al., 2010; Leithwood et al., 2004; Leithwood & Riehl, 2005; Sebastian et al., 2017).

Though there is a relationship between principal leadership and student achievement and school climate, there is debate about how direct that relationship is and therefore, how to measure such a complex relationship (Blase, 1987; Brown & Wynn, 2009; City, Elmore, Fiarman, & Teitel, 2009; Greenlee & Brown Jr., 2009; Johnson et al., 2012; Leithwood et al., 2004; Leithwood & Riehl, 2005). Some research shows that principals contribute to student achievement *indirectly* through their actions that influence school and classroom conditions, which poses a problem as to how to create a direct link to student achievement (Hallinger, 2003, 2005; Hallinger & Heck, 1996; Southworth, 2002). Other theories show principal leadership has significant *direct* and *indirect* effects on student achievement and school climate, particularly with school-restructuring initiatives and teacher-perceived student outcomes (Leithwood, 1994). Other studies argue that improved student achievement and school culture due to a principal's leadership may depend on how the principal adapts their leadership to the context (Halverson & Kelley, 2017; Leithwood & Sun, 2012; Orr, 2013; V. M. Robinson et al., 2008; Urick & Bowers, 2014b). However, even though *researchers* acknowledge

the difficulty in measuring the effects of principal leadership, in the American public school system, states are mandated to evaluate principals through measures that are statistically unsound and fail to acknowledge the adaptive elements of leadership due to context (Close et al., 2018; Day et al., 2016; Halverson & Kelley, 2017; Leithwood & Riehl, 2005; Leithwood & Sun, 2012; Orr, 2013; V. M. Robinson et al., 2008; Sun & Youngs, 2009). Hence, while principal leadership is significant, current methods are insufficient in assessing changes due to context.

Though principal evaluations are one of the primary vehicles for assessing principal leadership, there are four main reasons they are failing to do so adequately (Blase, 1987; Brown & Wynn, 2009; Greenlee & Brown, Jr., 2009; Johnson et al., 2012; Leithwood et al., 2004). First, if principal evaluation systems start with the premise that if the role of the principal is significant, then their assessment and development should also be significant (Drago-Severson, 2012; Fenton, 2016; Kegan & Lahey, 2016). However, principal evaluation models often fail to acknowledge and evaluate the intricacy of the principal leadership role (Day et al., 2016; Halverson & Kelley, 2017; Leithwood & Riehl, 2005; Leithwood & Sun, 2012; V. M. Robinson et al., 2008). Second, many current models use quantitative measures of principal leadership that are statistically inaccurate in measuring student and school academic growth (Close et al., 2018; Fuller et al., 2015; Piro et al., 2011). Third, there is often a disconnect between principals' perceptions of district expectations and goals for principals and what aspects of a principal's leadership are measured through the evaluation (Close et al., 2018; Ehlert et al., 2016; Orr, 2013; Sun & Youngs, 2009). Principals communicate that much of the feedback received was not impactful on their practice (Sun & Youngs, 2009; Tran &

Bon, 2015). Fourth, principal evaluation systems are driven by general federal and state policies, but the more systems align with instructional leadership activities at the local level, the more closely they measure principal leadership (Fenton, 2016; Hallinger et al., 2014; Sun & Youngs, 2009). In summary, while principal evaluations could be a powerful opportunity to measure principal leadership and support principal growth because of the way they are currently structured, they often do not (Corcoran et al., 2013; Grissom, Kalogrides, & Loeb, 2015; Sun & Youngs, 2009; Tran & Bon, 2015).

How Is Principal Leadership Defined?

The role of the principal is historically linked to the supervision of instruction and student achievement, and from the start, the relationship has been challenging to measure with much debate about which leadership practices lead to the best results (Derrington, 2018; Hallinger, 2011, 2018b; Leithwood & Riehl, 2005). When principal leadership research first appears in the1920's, principals followed a top-down model of supervision (Derrington, 2018). However, twenty years later, the principal supervision model switched to become more collaborative with the duties of a principal expanding to include facilitating group learning and shared decision-making (Derrington, 2018). Teacher performance ratings first appear in the 1950s to provide differentiated support, particularly around specific skills, such as spelling or arithmetic (Derrington, 2018). In the effective schools era of the late 1980s, "strong instructional leadership" is identified as a critical factor for effective urban elementary schools, causing a focus on the relationship between principals and instruction in a way that was not previously studied (Derrington, 2018; Hallinger, 2005, 2011, 2018b; Philip Hallinger & Joseph Murphy,

1985; Hitt & Tucker, 2016; Leithwood & Riehl, 2005; Prestine & Nelson, 2005; Printy, 2008). While this met a positive reception from policymakers, skeptics questioned how replicable the model was to all school principals and whether the model was merely meeting the public's expectations of a principal rather than being research-based (Hallinger, 2011; Leithwood, 1994). Then, during the change-oriented educational policy environment of the 1990s that focused on restructuring and transformation to meet the challenges of twenty-first-century school requirements, researchers began to study transformational leadership (Berkovich, 2016; Leithwood, 1994). While studies indicated the positive effect that transformational leadership has on school culture, the link to improving student achievement was not clear (Bass & Avolio, 1993; Bogler, 2001; Burns, 1978; Derrington, 2018; Geijsel, Sleegers, Leithwood, & Jantzi, 2003; Leithwood & Jantzi, 2005, 2008; Leithwood et al., 1998; Nguni et al., 2006). Research then built on aspects of both instructional and transformational leadership to create shared or integrated instructional leadership, which again showed a correlation with positive school culture and organizational learning with a nuanced link to student achievement (Bryk et al., 2010; Hitt & Tucker, 2016; Marks & Printy, 2003; Prestine & Nelson, 2005; Printy et al., 2009; Reich, 1988; Spillane, Halverson, & Diamond, 2001; Spillane et al., 2004). In recent years, the debate between which theories of leadership are correct has subsided as researchers have come to focus on the specific leadership skills from all three models needed by principals to affect change (Boyce & Bowers, 2018b; Cosner & Jones, 2016; Hallinger, 2003, 2011; Leithwood & Sun, 2012; Marks & Printy, 2003; Printy et al., 2009; V. M. Robinson et al., 2008; Thoonen, Sleegers, Oort, Peetsma, & Geijsel, 2011). The consistent patterns of impact on student achievement across the three theories

(instructional, transformational, shared) has led to the development of a leadership for learning framework, which combines forty years of research to highlight four key dimensions of leadership for student improvement (Hallinger, 2011).

Over the past thirty years, the educational research literature has shifted from traditional theories of leadership to integrated theories of leadership that highlight key actions and practices of principals who affect student and teacher learning (Cosner & Jones, 2016; Day et al., 2016; Hallinger, 2011; Leithwood & Jantzi, 2008; Leithwood et al., 1998; Nguni et al., 2006; Spillane et al., 2004). The three main traditional theories, as identified above, that have dominated the research around principal leadership are instructional leadership, transformational leadership, and shared or integrated instructional leadership. As current models of principal leadership are based on these three theories, I will discuss each of them in detail below. This section will conclude with current leadership theories.

Because the link between principal leadership and student achievement is difficult to measure, the question researched is often not merely "does principal leadership have an effect" but "to what extent" (Bryk et al., 2010; Hitt & Tucker, 2016; Marks & Printy, 2003; Prestine & Nelson, 2005; Printy et al., 2009; Reich, 1988; Spillane et al., 2001, 2004). This type of research analyzes the particular effectiveness of principal leadership with an effect measure. Results are often reported in effect sizes, which are generally expressed as the standardized mean difference between an experimental and control group (Hendriks & Scheerens, 2013). While effect sizes will vary in significance depending on the type of statistical analysis run, in educational research generally an effect of 0.20 is considered small, 0.40 is considered moderate, and 0.60 is considered

large (Cohen, 1992; Hattie, 2009). Thus, the reported significance of effect sizes is vital to the results of the principal leadership research that follows.

Instructional Leadership

After forty years of research, a preponderance of evidence from research on instructional leadership indicates that principals make a *small* and *indirect* yet significant contribution to student achievement outcomes, particularly through their actions that impact school and classroom conditions (Hallinger, 2003, 2005, 2011, 2018b; Hallinger, Bickman, & Davis, 1996; Southworth, 2002; Urick, Wilson, Ford, Frick, & Wronowski, 2018). Proponents of instructional leadership argue that the most substantial effects of the instructional leader are shaping the school's mission and direction, managing the instructional program, and promoting a positive school climate. Still, the effect size of this appears to vary by school level, and the socioeconomic status of the students as context dramatically influences the type of instructional leadership the principal delivers (Hallinger, 2005, 2011, 2018a, 2018b; Urick, 2016; Urick & Bowers, 2014b; Urick et al., 2018). Although that context is recognized as such a significant factor, few studies have taken it into account when studying instructional leadership, causing some of these methods to be flawed (Hallinger, 2018). Similarly problematic is that instructional leadership is highly correlated with organizational management, begging the question as to whether even indirect significant effects are due to leadership or merely management (Sebastian, Allensworth, Wiedermann, Hochbein, & Cunningham, 2018). Additionally, the evidence shows a connection between effective instructional leaders and school climate, a key aspect of principal leadership, is moderate (Boyce & Bowers, 2018a; Brown & Wynn, 2009; Greenlee & Brown Jr., 2009; Jason A. Grissom & Susanna Loeb,

2011; Johnson et al., 2012; Southworth, 2002). Even though researchers argue for the importance of instructional leadership, there is debate as to its effectiveness.

While instructional leadership aims to affect student achievement, research is mixed as to whether it is an effective framework for principals as much of these effects are small and indirect (Boyce & Bowers, 2018b; Hallinger, 2003, 2005, 2011, 2018b; Philip Hallinger & Joseph Murphy, 1985; Hitt & Tucker, 2016). First, a key component of instructional leadership is managing the instructional program, which includes supervising and evaluating instruction and creating an image of instructional leaders displaying hands-on involvement in the classroom (Hallinger, 2005). However, the preponderance of evidence on instructional leadership shows that principals' effects on the classroom happen not through supervision of instruction, but by modeling and through the school culture (Boyce & Bowers, 2018b; Hallinger, 2005, 2011, 2018b). Secondly, instructional leadership seems influenced by context, which changed the effect size (Hallinger, 2003, 2005, 2018b; Philip Hallinger & Joseph Murphy, 1985; Southworth, 2002; Urick & Bowers, 2014b). For instance, there is some evidence that classroom visits have a positive effect on elementary schools. Still, in high schools, teachers often have more content expertise in their particular content area than principals, making managing the instructional program more difficult for principals to lead (Hallinger, 2005, 2018). Instructional leadership also appears to be too narrow a definition of principal leadership, as it does not encompass all that a leader does, individually as managers and politicians (Hallinger, 2003, 2018b; Leithwood & Riehl, 2005; Reich, 1988). As a research concept, instructional leadership is general with little evidence of the specific practices that make for a successful instructional leader and of

how much time is needed to spend on those behaviors to affect student achievement (Boyce & Bowers, 2018b; Philip Hallinger & J Murphy, 1985; Prestine & Nelson, 2005). Additionally, how teachers perceive instructional leadership may be different than how principals perceive it, begging the question of whether or not instructional leadership is effective if it is not recognized in the leader by its followers (Bowers et al., 2017; Boyce & Bowers, 2018a; Reich, 1988; Urick & Bowers, 2019). Though instructional leadership may have important aspects, it is difficult to provide a research-based tie to its effectiveness in improving student achievement.

Transformational Leadership

In transformational leadership, the purposes of the leader and the followers become fused so that they are working together towards a common cause (Burns, 1978). In his seminal work, *Leadership*, Burns describes leadership as the opposite of power holding and posits that a transformational leader moves their followers to a place of higher morality. He contrasts this with transactional leadership, a managerial style of leadership in which the leader and the follower exchange goods with no enduring purpose. Bass (1985) builds on this theory by proposing that the measure of transformational leadership is the amount of influence a leader has over their followers. Bass and Avolio (1994) create a model for this measurement that involves four factors: individualized consideration, intellectual stimulation, inspirational leadership, and idealized influence. Transformational leadership aims to literally transform the leader and the followers with the end goals of greater justice and equality (Bass, 1985; Bass & Avolio, 1994; Burns, 1978; Printy et al., 2009).

Transformational leadership is developed more specifically for principals through the work of Leithwood and his team (Leithwood, 1994; Leithwood & Jantzi, 2008; Leithwood et al., 1998). His theory rests on the fact that while instructional leadership was the primary mode of leadership for the past decade, a different style of leadership was needed to meet the challenges of motivating and moving staff through the large scale reform efforts of the 1990s and early 2000s (Berkovich, 2016; Leithwood, 1994). Furthermore, he noted that most reform efforts would occur in secondary schools, which would be too large for instructional leadership to be successful. In testing his framework through survey results from teachers and administrators working on the restructuring of schools, he found that transformational leadership had significant direct and indirect effects on progress with school-restructuring initiatives and teacher-perceived student outcomes with regression coefficients in the 0.60 to 0.70 range (Leithwood, 1994). Leithwood's findings support Burns's (1978) definitions of transactional and transformational leadership. Still, he argues that while transformational leadership is value-added as Avolio and Bass (1988) suggest, the base point may be individual consideration and not transactional leadership. Leithwood questions whether this type of leadership can be adequately measured with specific practices as there were high levels of critical thinking in these leaders that could not always be distinguished through their behavior.

Despite the complexity of measuring the effects of transformational leadership, further studies attempt to do so by building on Leithwood's work and defining specific characteristics of transformational leaders (Leithwood & Jantzi, 2008; Leithwood et al., 1998; Leithwood & Sun, 2012; Nguni et al., 2006; Thoonen et al., 2011). Three main

aspects of transformational leadership, as demonstrated through principals, were identified as charismatic leadership, intellectual stimulation, and individualized consideration (Nguni et al., 2006). Transformational leadership was found to contribute to organizational learning and learning processes significantly and, in particular, is positively associated with innovation in a school climate (Leithwood et al., 1998; Moolenaar, Daly, & Sleegers, 2010; Thoonen et al., 2011). It has moderate, positive, and significant effects on articulating a vision or shared goals (.67), working environment (.56), providing individualized supports and improved instruction (.55) with moderate effects on positive organizational culture (.44) and shared decision-making (.36) (Leithwood et al., 1998; Leithwood & Sun, 2012; Thoonen et al., 2011). Survey findings provide evidence that teachers prefer to work with a transformational leader over a transactional one in part because they often foster a work-related closeness that helps to build the school climate (Bogler, 2001; Moolenaar et al., 2010). Studies on transformational leadership reveal that it has moderate effect sizes on the organization of the school environment.

Despite these findings, research is inconclusive on exactly how transformational leadership supports specific leadership practices that impact teaching practices and student achievement (Thoonen et al., 2011). One concern is the loose definition of transformational leadership, which can be almost tautological, as there is some thought that the definition has broadened over time to include those activities, which proved successful for leaders (Berkovich, 2016). Also problematic is that while transformational leadership was heralded as a solution for high schools, many of the studies that showed positive effects were conducted in primary schools. There is some evidence that the

positive effects of transformational leadership are muted in larger districts and secondary schools (Leithwood & Jantzi, 2008; Nguni et al., 2006; Thoonen et al., 2011). Similarly, research suggests that transformational leadership is less likely to be enacted in specific contexts, such as when strong accountability measures are adopted (Cosner & Jones, 2016). Additionally, evidence of transformational leadership having a positive effect on job satisfaction and organizational commitment is weak (Bolger, 2001; Nguni et al., 2006). Finally, in studying the relationship that transformational leadership has on student achievement, there were significant positive effects, but the effect sizes were weak (0.28, 0.24, and 0.25) (Leithwood & Jantzi, 2008). Though transformational leadership was critical during the reform era, with its inconclusive evidence on effecting teacher and student achievement, it may not have had as essential a place in the current achievement era (Boyce & Bowers, 2018b; Hallinger, 2005; Leithwood, 1994; Leithwood et al., 1998; Thoonen et al., 2011; Urick et al., 2018).

Shared Instructional Leadership

The power of transformational leadership lies in the ability to move followers forward towards a higher purpose by involving them in the work (Bass, 1985; Burns, 1978). However, early models measured the leadership of the principal alone (Hallinger, 2003; Leithwood & Sun, 2012; Nguni et al., 2006; Thoonen et al., 2011). Shared instructional leadership is born out of the theory that authentic leadership involves and develops followers so that they also join in the leadership work (Marks & Printy, 2003; Printy et al., 2009). It incorporates the instructional leadership finding that the job of leading and developing instruction is too complex to put on any one person (Hallinger, 2003, 2005). Shared instructional leadership defines educational leadership as a

"synergistic power of leadership shared by individuals through the school organization" (Marks & Printy, 2003). The active collaboration of the principal and their followers, often teachers, on curriculum, instruction, and assessment occurs informally and formally with specific roles (Marks & Printy, 2003; Printy et al., 2009). The thinking and behavior of the leader need to be considered together in the interaction of the leader and the followers in that specific context (Spillane et al., 2004; Urick & Bowers, 2014). In shared leadership, principals play a dual role in leading the community while being a part of it (Boyce & Bowers, 2018b; Byrne-Jiménez & Orr, 2012).

There is a growing body of research on the impact of shared instructional leadership. It was first researched by Marks and Printy (2003) by using a hierarchical linear model to study twenty-four restructured schools. While the study was small and may not be generalizable, they found that transformational and shared instructional leadership were found together in a school, there is a substantial impact on school performance as measured through pedagogy and student achievement (Marks & Printy, 2003). Transformational leadership was seen as a prerequisite for shared instructional leadership as principals were found to be critical to the development of leadership in teachers (Marks & Printy, 2003). Printy, Marks, and Bowers (2009) built on this research by studying the three schools, which exercised the highest levels of both transformational and shared leadership in the previous study. Though a small sample size, they find that teachers and principals worked together to initiate school improvement efforts around a strong belief that all students could learn. While principals buffered teachers from managerial tasks, they shared tasks around curriculum and instruction, such as hiring new teachers, data teams, and designing instructional strategies. Research on successful

shared instructional leadership reveals that it builds on both instructional and transformational leadership.

The research on shared instructional leadership, once again, proves leadership to be complicated; however, it does have the largest effect on student academic growth (Robinson et al., 2008; Urick & Bowers, 2014). Shared instructional leadership is highly correlated with interpersonal synergies in the principal and seen as a strategy to prevent principal burnout because the work is shared between the leader and the followers (Donaldson, 2001; Marks & Printy, 2003; Printy et al., 2009). The development of leadership or shared instructional leadership in teachers leads to teacher empowerment, which increases teacher satisfaction and teacher commitment to their job (Day et al., 2016; Marks & Louis, 1999; Muijs & Harris, 2003; Printy et al., 2009; Ware & Kitsantas, 2007; Zembylas & Papanastasiou, 2005). While shared instructional leadership is seen to be a necessary prerequisite to improve instructional quality and academic outcomes, researchers acknowledge that this relationship is complicated (Jacob, Goddard, Kim, Miller, & Goddard, 2015; Printy et al., 2009; Sebastian, Allensworth, & Stevens, 2014).

Principal Leadership Typology

Research has historically tied principal leadership to specific theories and models of leadership, such as instructional, transformation, and shared instructional leadership (Bowers et al., 2017; Derrington, 2018; Hallinger, 2005, 2011, 2018b; Philip Hallinger & Joseph Murphy, 1985; Hitt & Tucker, 2016; Leithwood & Riehl, 2005; Marks & Printy, 2003; Prestine & Nelson, 2005; Printy, 2008; Printy et al., 2009; Urick & Bowers, 2014b). However, those styles have significant overlap and do not fully encapsulate the complexity of principal leadership (Blase, 1987; City et al., 2009; Day et al., 2016;

Greenlee & Brown Jr., 2009; Johnson et al., 2012; Leithwood et al., 2004; Leithwood & Riehl, 2005). A few recent studies have examined leadership types to examine how leaders enact leadership and what these models look like in practice (Bowers et al., 2017; Urick & Bowers, 2014b). Other studies build on the idea that educational leadership is layered because it changes over time based on the needs of the school (Day et al., 2016).

In their innovative Latent Class Analysis (LCA), Urick and Bowers (2014) show that leadership follows more of a typology than a specific style due to contextual factors, such as school size, region, accountability performance, and principal background. Using 7,650 public schools and principals from the 1999 Schools and Staffing Survey, they ran a LCA that included school context and survey responses to items in regards to transformational leadership, shared instructional leadership, and management. Their study revealed three types of leaders: integrating (high transformational leadership and high shared instructional leadership), balkanizing (low transformational leadership and low shared instructional leadership), and controlling (high transformational leadership and low shared instructional leadership). Significant to these findings were the three distinct types that did not follow any specific leadership styles but merged leadership styles to enact particular behaviors due to the context. For example, balkanizing principals were more often male in small, rural schools. This finding illustrates that leadership may be more multifaceted than one particular style as it is adaptive based on context, which aids in our understanding of how principals enact specific leadership practices as opposed to overarching styles.

Further research examines how principals often combine elements of instructional and transformational leadership to create improvement in a school over time based on the

specific needs and context of the school. In a prize-winning, three-year, mixed-methods study, Day, Gu, and Sammons (2016) use national survey data, student assessment scores, and ten multiyear case studies to discover that both transformational and instructional leadership were needed to sustain improvement. They found that principals who had improved academic outcomes were able to respond to recognize the needs and motivations of others and react accordingly to the context. They use the term "layered leadership" to describe the process by which the principal changes their style of leadership, using aspects of both instructional and transformational leadership to adapt to the school's needs during their tenure. Day, Gu, and Sammons conclude that an educational leader's ability to sustain improvement over time is not based on a particular leadership style, but rather their ability to demonstrate specific leadership traits and behaviors at the appropriate time as determined by context.

As we consider the complex nature of leadership, both studies have substantial implications for research and practice (Leithwood & Riehl, 2005). If leadership is more tied to a typology and changes with the context or the time spent in the position, then less time needs to be spent researching overarching leadership theories such as transformational and instructional leadership, and more time needs to focus on specific leadership behaviors (Hallinger, 2018). Similarly, this research challenges the notion that one style of leadership will produce the most significant effect on student achievement. It introduces the idea that leadership is more tied to a context that a set style (Day et al., 2016; Hallinger, 2018; Halverson & Kelly, 2017; Robinson et al., 2008; Urick & Bowers, 2014).

Principal Leadership: What does Research Recommend?

After decades of research on instructional, transformational leadership, and shared instructional leadership, researchers found that there were more similarities than differences (Berkovich, 2016; Boyce & Bowers, 2018b; Hallinger, 2003, 2011, 2018b; Leithwood & Sun, 2012; Marks & Printy, 2003; Printy et al., 2009; V. M. Robinson et al., 2008; Thoonen et al., 2011; Urick & Bowers, 2014b). For example, shaping the vision or mission for the school was a critical factor in both frameworks (Hallinger, 2003, 2005, 2011, 2018; Leithwood et al., 1998; Leithwood & Sun, 2012; Marks & Printy, 2003). When principals are exhibiting instructional leadership, transformational leadership, or shared instructional leadership, there is a focus on shaping and developing the school culture (Bogler, 2001; Boyce & Bowers, 2018b; Burns, 1978; Hallinger, 2003, 2005, 2011, 2018b; Leithwood et al., 1998; Leithwood & Sun, 2012; Moolenaar et al., 2010; Muijs & Harris, 2003; Printy et al., 2009; Thoonen et al., 2011; Ware & Kitsantas, 2007). Recent studies of instructional leadership show that teachers collectively view leadership differently than they do as individuals, affirming the belief that instructional leadership is done at the collective level, an idea that would be formerly associated with transformational leadership (Boyce & Bowers, 2018a). In fact, several studies argue that transformational and instructional leadership are linked (Boyce & Bowers, 2018b; Day et al., 2016; Marks & Printy, 2003). Some research suggests that transformational leadership may need to be present for instructional leadership to occur and that both styles of leadership are necessary for success (Day et al., 2016; Marks & Printy, 2003).

Despite the similarities in the research, neither theory provides a strong link to student achievement. In their 2008 study, Robinson, Lloyd, and Rowe used two meta-

analyses to compare the effects of transformational leadership and instructional leadership on student outcomes. In the first analysis, they analyzed 22 studies to examine the effects of transformational and instructional leadership with the finding that the average effect of instructional leadership (0.42) was three to four times that of transformational leadership (0.10) and larger than other leadership styles (0.30) included in the study. Similarly, the effect size was larger for instructional leadership in participating and promoting teacher learning. This finding is in contrast to Leithwood and Jantzi's (2005) review of thirty-three studies, of which approximately half were found to show that transformational leadership had a small indirect influence on social student outcomes. Of course, transformational leadership does focus more on adults, which could contribute to this finding. Robinson, Lloyd, and Rowe's second meta-analysis examined specific leadership practices from both instructional and transformational leadership that had an impact on student achievement (Robinson et al., 2008; Robinson, 2011). This analysis showed strong average effects for promoting teacher learning and development (0.84) and moderate effects for establishing goals and expectations (0.42) and ensuring quality teaching (0.42). These leadership behaviors have larger effect sizes than either of the two leadership theories. The other two domains, resourcing strategically (0.31) and ensuring an orderly and safe environment (0.27), were between a small and moderate effect. Robinson, Lloyd, and Rowe's study contributes to the body of research on principal leadership that shows leadership styles are more integrated than previous research suggested (Cosner & Jones, 2016; Day et al., 2016; Hallinger, 2011; Leithwood & Jantzi, 2008; Leithwood et al., 1998; Nguni et al., 2006; Spillane et al., 2004; Urick & Bowers, 2014b). It calls for more research on specific individual principal leadership

behaviors that support teaching and student learning (Prestine & Nelson, 2005; V. M. Robinson et al., 2008; Stein & Spillane, 2005).

So where does that leave us? As Robinson, Lloyd, and Rowe (2008) suggest, since specific theories do not produce the academic gains in students that we might expect to see, it is helpful to identify commonalities in specific principal leadership behaviors. Research also points to the question of context as perhaps a more significant mitigating factor than any type of leadership theory (Day et al., 2016; Hallinger, 2005; Southworth, 2002; Urick & Bowers, 2014). Principal leadership may not fit one specific leadership framework but be a combination of leadership behaviors that the leader selects to respond appropriately to situations in different contexts (Day et al., 2016; Hallinger, 2005). Thus, the question is not under what theory do leaders best operate or have the greatest impact on student achievement, but what set of behaviors and skills allow principals to be responsive in the correct context (Day et al., 2016; Hallinger, 2018; Urick & Bowers, 2014). Principal leadership is multidimensional, and school context factors may predict the type of leadership that a principal shows in a particular school (Urick & Bowers, 2014).

Researchers seem to agree that there are two ways in which leaders influence teachers and students. First, the leader is *socially* distributed across the organization, meaning that the work is not done by the principal alone, but through his or her orchestration of different key actors who contribute collaboratively to the achievement of school goals (Halverson & Kelley, 2017; Leithwood & Jantzi, 2008; Marks & Printy, 2003; Printy et al., 2009; Spillane, 2006; Spillane et al., 2004; Waters, Marzano, & McNulty, 2003). Second, the leader is *situationally* distributed across the organization,

creating or gathering different resources according to the context of the school (Halverson & Kelly, 2017; Leithwood, 1994; Spillane, 2006; Spillane et al., 2004; Urick & Bowers, 2014). Specific practices can be measured to identify the impact of principals and to learn how they adapt to different contexts. However, more research needs to be done on this topic (Halverson & Kelly, 2017; Leithwood & Sun, 2012; Robinson et al., 2008).

Assessing Principal Leadership: Current Policies and Methods

In addition to the question of what theories and aspects of principal leadership are most important for school culture and student achievement, there is the question of how to measure a complex task such as principal leadership. After No Child Left Behind (NCLB), evaluation systems are dominated by accountability with standardized test scores for students that tie to evaluations for teachers and principals (Henig, 2013). While previous eras used scores as formative measures for budget, curriculum, and staffing decisions, the era of accountability uses these scores as measures for both resources and corrective labels and measures (Henig, 2013; Horsford, Scott, & Anderson, 2018). This market shift to become more results-oriented has created a policy that uses teacher and principal evaluation as the primary vehicle for professional development and growth (Riley & Meredith, 2017). However, research has shown that there is often a disconnect between principal evaluations and learning-centered behaviors (Close et al., 2018; Ehlert et al., 2016; Orr, 2013; Sun & Youngs, 2009). Evaluation models are built on a Neo-liberal approach that views education as a free market, where teachers and principals need to be rated so that the consumers, that is parents and students, can make a decision

based on the “facts” of the data (Horsford et al., 2018). Sadly, the data used through Value-added Models is statistically unsound, making the comparison biased, particularly against low performing schools (Close et al., 2018; Jason A. Grissom & Susanna Loeb, 2011; Orr, 2013). Additionally, they often fail to take into account the context of the principal leadership, including how the era of accountability has affected their leadership (Cosner & Jones, 2016; Hallinger, 2018). Many current evaluation models examine principals generally and fail to acknowledge the situationally and socially distributed nature of principal leadership (Halverson & Kelly, 2017; Leithwood & Jantzi, 2008; Marks & Printy, 2003; Printy et al., 2009; Spillane, 2006; Spillane et al., 2004).

How Public Policy has Shaped Principal Evaluation Methods:

Basis of Policy Context

Federal and state legislative policy has dictated the extent to which high stakes testing is used for accountability measures. This reform narrative originated in the Reagan era with the publishing of the Nation at Risk report (Derrington, 2018; Horsford et al., 2018). Rather than affirm previous values of equity and efficiency, it framed the problem of education around standards, quality, and excellence with a top-down approach for principals supervising teachers (Derrington, 2018). In response to the narrative of failing schools, the Elementary and Secondary Education Act (ESEA) was reauthorized as No Child Left Behind, which gave increased power to the federal government and executive branch, holding schools accountable for student achievement according to subgroups (Derrington, 2018; Horsford et al., 2018). Federal money followed through the Race to the Top grant solidified their use for accountability in

requiring assessment scores to be used for evaluation of teacher and principal effectiveness (Close et al., 2018; Derrington, 2018; Orr, 2013). This funding left urban districts most vulnerable as their scarce financial resources often left them with little alternative but to adopt the federal policy (Hunter & Donahoo, 2003). While the legislation pursued accountability in the name of equity, the stratification that has occurred along racial lines has only furthered inequity by normalizing a narrative of under-performing Latino and Black subgroups and reinforcing the narrative of failing urban schools where the concentrations of those groups of students are the highest (Horsford et al., 2018). Although the narrative framed the problem as failing schools and the solution as accountability, the problem was equity, and these initiatives served only to highlight and increase gaps (Horsford et al., 2018).

In compliance with the mandates, states used three different accountability systems to measure the student growth that would be attached to principal and teacher evaluations. In addition to a stronger belief in a market-based system, these performance metrics arose because of research showing that schools and teachers differ significantly in terms of student growth and because of the difficulty in linking achievement to observable characteristics (Ehlert et al., 2016). The first of the methods is a student growth percentile (SGP), which are student-level conditional performance percentiles relative to a peer group that controls for covariates in the school environment (Ehlert et al., 2016). The most common measure is the use of Value-added Measures (VAMs), as 46 states use either a one-step or two-step model (Close et al., 2018). When comparing all three methods, as Ehlert et al. (2016) did in their study, the two-step fixed effects VAM

has the most validity as it facilitates context-conditional comparisons. The first two methods (SGP and one-step VAM) favor advantaged schools.

Despite evidence that the two-step VAMs would account for poverty effects, all three methods have significant limitations. Teacher and principal evaluation systems that are based on SGPs or one-step VAM estimates will disproportionately identify schools as underperforming (Ehlert et al., 2016). As noted by leading educational organizations, such as the American Educational Research Association (AERA) and the American Statistical Association (ASA), two-step VAMs are fraught with mathematical issues and are unreliable, invalid, not transparent, unfair, and contain measurement errors (Bowers & White, 2014; Close et al., 2018; Jason A. Grissom & Susanna Loeb, 2011; Orr, 2013).

Policy experts remain skeptical of hailing value-added models for principals as effective measures due to their limited scope and the indirect effect of principal leadership (Jason A. Grissom & Susanna Loeb, 2011; Leithwood et al., 2004). While the studies on principal evaluation systems did show that principals have a positive effect on achievement and that the size of the influence varies by poverty rate, the exact effect size is difficult to measure (Orr, 2013). Value-added models are limited in their ability to assess principals. VAM's fail to recognize the effects that a prior principal might have had on the school and the reality that new principals are often assigned to the most challenging schools where there is the highest turnover. Other studies found the covariates in the VAM's "so noisy" that they might not be useful in practice. Their utility is highly problematic, given their violation of the central assumptions of statistics (Bowers & White, 2014). The inconsistencies are so great that it behooves policymakers

to question whether they should be used since they are not fair or unbiased (Grissom et al., 2015).

Evaluation Methods: Current Policy Context

The recent authorization of the Every Student Succeeds Act (ESSA) granted states the ability to set their own accountability goals within the federal framework; however, few states have changed their policies or their principal evaluation methods. They continue to make high stakes personnel decisions, such as continued employment and salary changes using inaccurate methods (Close et al., 2018; Corcoran et al., 2013; Derrington, 2018; Fuller et al., 2015; Piro et al., 2011; Sun & Youngs, 2009). This lack of congruence between policy adoption and fidelity in implementation may have an overall negative effect on principal supply and quality as many of these policies operate under the assumption that the labor market is equal. However, in education, the market is not equal. Consequently, not all schools can recruit the highest-quality teachers and principals (Close et al., 2018; Ehlert et al., 2016; Fuller et al., 2015).

While ESSA has granted flexibility to states and some have exercised that flexibility, most states continue to use flawed value added measures. Forty-five states continue to give performance ratings using letter grades and scores from 0-100 (Close et al., 2018; Piro et al., 2011; White et al., 2012). These scores often negatively label urban schools, which frequently receive lower scores or are lower on the ranking (Fuller et al., 2015; White et al., 2012). All states contain student growth components as part of their principal evaluation system (Fuller et al., 2015). These student growth components take five different forms: Value Added Model (controls), Value Added Model (simple), Student Growth Percentile (SGP), Student Learning Objective (SLO), and Growth to

Proficiency. While none of these models have a great degree of statistical accuracy, some do have more accuracy than others (Bowers & White, 2014; Fuller et al., 2015; Goldring et al., 2008; Liu et al., 2015; Shoaf et al., 2013). Therefore, principals, regardless of their effectiveness, do not have an equal opportunity to be rated as effective across the United States.

Likewise, while all of the states contain student academic growth measures and a rubric piece to their principal evaluation systems, the weighting of the student academic growth measures vary. Half of the states do not specify a state percentage, perhaps leaving it up to the district and local bargaining units to decide how much student achievement should be weighted in a principal evaluation (Fuller et al., 2015). Twenty-four percent of states weight student academic growth measures at 50%, about eighteen percent of states weight it between 20% and 49%, and eight percent of states weight it less than 20%. Sixty-nine percent of states use other student academic outcomes, such as Student Learning Objectives, as an additional part of the principal evaluation (Fuller et al., 2015). Only 28% of states use an assessment of school climate or working conditions as a part of their principal evaluation. Not even twenty percent of states use metrics on teacher quality and retention as part of the principal evaluation system.

Continued implementation of principal evaluation systems is similarly weak, with 35% of states saying that there was state oversight, only 24% received training, and only 16% had developed a process for principals to correct mistakes (Fuller et al., 2015). In California and New Jersey, studies show as high as 99% of principals are rated at the highest level, calling into question the validity of the method (Herrmann & Ross, 2016; White et al., 2012). Despite regulation changes through ESSA, principal evaluation

methods remain the same, particularly in the area of high stakes test results for accountability, making the argument that policy change could be "too little, too late" as the retraction is not as easy as the creation (Henig, 2013).

Of course, the flexibility of ESSA is new, so states can put in processes to continue to revise and change principal evaluation methods. However, current research seems to indicate that an improvement process is only in place for less than half of the states. Only 40% of states have a system in place to evaluate the evaluation system at any point in spite of research that indicates the importance of doing so (Fuller et al., 2015; Gullickson & Howard, 2009). In a recent survey of representatives from twenty-eight states, only 25% of survey participants reported that revising current principal evaluation systems was currently a priority (Fuller et al., 2015; Riley & Meredith, 2017). Ohio, Delaware, and Arkansas are a few of the states that are revising their principal evaluations systems to be more effective. Kentucky, Washington, Arizona, Wisconsin, and Idaho have chosen to revise principal evaluation systems by developing stronger systems to train principal supervisors on how to correctly evaluate principals (Riley & Meredith, 2017). Other states, such as New Jersey and South Carolina, are moving away from an accountability process to systems of principal growth that include self-reflection, authentic evidence, and discussions about practice. Many districts are concerned about the amount of principal turnover and want to use evaluation as a means of support and not accountability (Turnbull, Anderson, Riley, MacFarlane, & Aladjem, 2016). States are trying different methods of improvement because this implementation is new. There are few state models and examples of an effective principal evaluation system. Researchers suggest states collaborating on collecting resources and conducting case studies to

provide a stronger knowledge base of effective principal evaluations (Riley & Meredith, 2017). Case studies and examples of stronger models would allow states to improve their evaluation systems in authentic ways that could move away from inequitable, statistically inaccurate value based measures.

Alternative Evaluation Methods: The Brave Few

Due to concerns regarding the validity of current principal evaluation systems, many researchers are urging for new principal evaluation tools that can be used as an accurate measure of principal leadership and support a cycle of continuous improvement (Fenton, 2016)). New principal evaluation methods include the creation of learning goals, coaching, and the use of 360-degree feedback (Corcoran et al., 2013; Goldring et al., 2008; Liu et al., 2015; Shoaf et al., 2013; Sinnema & Robinson, 2012; Tran & Bon, 2015; Williams, 2015). Many of these methods incorporate both the situationally and the socially distributed nature of leadership (Halverson & Kelly, 2017; Leithwood, 1994; Leithwood & Jantzi, 2008; Marks & Printy, 2003; Printy et al., 2009; Spillane, 2006; Spillane et al., 2004). When critiqued for effectiveness, if the systems were more aligned with instructional leadership activities, the measurements of leadership were more reliable (Fenton, 2016; Sun & Youngs, 2009).

Standards or goals-based. One method is that of aligning principal evaluations to standards-based goals that are measured on the specific leadership context (Corcoran et al., 2013; Shoaf et al., 2013; Sinnema & Robinson, 2012; Tran & Bon, 2015; Williams, 2015). Principals were more committed to the standards-based goals that they set which were aligned with their practice and the specific needs of their district (Corcoran et al., 2013; Sinnema & Robinson, 2012; Sun & Youngs, 2009; Turnbull et al., 2016). Often

these goals were created with a principal supervisor who visited them regularly over the year to evaluate different aspects of their practice and provide mentoring and resources to enable the principals to meet their established goals (Corcoran et al., 2013; Turnbull et al., 2016). Other methods using goals contain a formative piece that the principals complete over time and a summative piece, which included an improvement plan (Shoaf et al., 2013). The formation of clear goals allows for a consistent standard in the principal evaluation model while enabling personalization based on context, something that most current models neglect (Fuller et al., 2015; Williams, 2015; White et al., 2012). Principals found this type of method to be fair, unbiased, and informative to their practice (Corcoran et al., 2013; Turnbull et al., 2016).

360-degree feedback. Other new principal evaluation models include the use of 360-degree feedback, which includes surveys from key stakeholders, such as parents, teachers, and staff (Goldring et al., 2015; Liu et al., 2014; Shoaf et al., 2013; Tran & Bon, 2015). Adding student and teacher surveys on school conditions to the evaluation model was found to strengthen the relationship between principal's evaluation results and the schools' value-added achievement gains in math and composite scores in both math and reading. The use of surveys was mainly incremental in the areas of instructional leadership and the classroom environment in informing how principals shape the environment for teachers (Liu et al., 2014). Qualitative studies on the method found that principals benefited from the transparency, feedback, and narrative as opposed to the previous method of weighted scores, as principals needed differentiation as they have different levels of experience (Goldring et al., 2015; Tran & Bon, 2015).

While 360-degree feedback, like all of the new models, needs training and expressly set goals to be effective, it accounts for the socially distributed nature of leadership that other methods neglect (Bowers et al., 2017; Halverson, 2010; Leithwood & Jantzi, 2008; Marks & Printy, 2003; Printy et al., 2009; Spillane, 2006; Spillane et al., 2004). Principal leadership strongly affects the work environment for teachers, having an impact on school climate and teacher retention (Blase, 1987; Brown & Wynn, 2009; Greenlee & Brown Jr., 2009; Johnson et al., 2012). Successful principals are systems thinkers who build communities of "collective learning" where teachers strive to solve problems together (Brown & Wynn, 2009; Griffith, 2004; Minarik, 2003). The need for principals to be system thinkers is particularly true in urban schools where principal support plays a critical role in teachers' decisions to stay specifically in a high needs schools (Greenlee & Brown Jr., 2009; Hughes, Matt, & O'Reilly, 2015; R. Ingersoll & May, 2016; Johnson et al., 2012). This critical aspect of principal leadership cannot be measured through VAMs or SGP, but only through feedback from the teachers themselves (Goldring et al., 2015; Liu et al., 2014; Shoaf et al., 2013; Tran & Bon, 2015).

Principal Evaluation as a Measurement for Leadership

While federal and state legislation plays a significant role in the type of principal evaluations that are conducted, policy requirements are not research-based, nor are they unbiased (Close et al., 2018; Ehlert et al., 2016; Grissom et al., 2012; Orr, 2013). In the post-ESSA context, states have the flexibility to create new plans, but few states are changing methods (Close et al., 2018; Fuller et al., 2015). Some states are attempting new methods of principal evaluation, such as revising current methods to align principal evaluation rubrics more closely to standards, the addition of coaching and support, and

expanding the role of principal supervisors. Early research seems to point to those methods as more effective (Corcoran et al., 2013; Fenton, 2016; Goldring et al., 2015; Liu et al., 2014; Shoaf et al., 2013; Sinnema & Robinson, 2012; Sun & Youngs, 2009; Turnbull et al., 2016; Tran & Bon, 2015; Williams, 2015). Nonetheless, continued research needs to be done around the success of each of these methods since even these newer models have barriers to successfully measuring principal leadership.

Principal leadership remains challenging to assess. Some argue that none of the approaches provide measures of leadership that are valid or reliable enough to be used for high stakes decisions (Fuller et al., 2015). Many research studies have design flaws that may influence data results. Researchers struggled to account for covariates fully, making it difficult to measure the direct impact of the principal leadership (Grissom et al., 2012; Shoaf et al.; Sun & Youngs, 2009). School populations involved in measures of impact studies were subject to a maturation threat, as they change over the analysis period (Fuller et al., 2015; Orr, 2013). Cohort data were not consistent as it did not account for transfers and registrations or higher populations of special education or economically disadvantaged students (Fuller et al., 2015; Orr, 2013). Similarly, several of the studies did not include teacher and student measures, which would provide a needed triangulation to the principal's response (Schreiber & Asner-Self, 2011; Sun & Youngs, 2009; White et al., 2012). Additionally, many of the studies on new methods, involve surveys and interviews from a small sample size, making them a design of low power (Corcoran et al., 2013; Ehlert et al., 2016; Liu et al., 2014; Schreiber & Asner-Self, 2011; Sun & Youngs, 2009; Turnbull et al., 2016; Tran & Bon, 2015). Theses validity threats affect the impact of the measurements.

There are also gaps in the type of data that is collected. Data on principal evaluation is often self-reported, causing a threat to the measurement validity since no data verification was completed (Close et al., 2018; Schreiber & Asner-Self, 2011; White et al., 2012). In many of the methods, terms for referring to principals were relative, such as “successful” and “high-rated” and as an evaluation is a human endeavor, inevitably interpretation of terms and implementation of methods varied (Orr, 2013; Shoaf et al., 2013). For studies that were evaluating methods that used standards-based goals, those goals were self-reported and not measured for complexity (Ehlert et al., 2016; Sinnema & Robinson, 2012). Lastly, there are substantial research gaps, as methods in charter and private schools are mainly unstudied, as they are not always required to use state-approved principal evaluation methods (White et al., 2012).

Principal Evaluation Systems: What Does Research Recommend?

Although much of what makes principal evaluations effective is yet to be comprehensively measured, researchers make several recommendations. First, researchers advocate for strong and straightforward principal evaluation tools that are explicitly aligned to professional learning standards which provide clarity and leadership expectations (Corcoran et al., 2013; Fenton, 2016; Micheaux & Parvin, 2018; Sinnema & Robinson, 2012; Tran & Bon, 2015; Williams, 2015). Second, rather than rely on assessments that use flawed measures of student accountability, principal evaluations should be aligned to district strategic goals and focus on instructional leadership activities (Close et al., 2018; Corcoran et al.; Leithwood et al., 2004; Sun & Youngs, 2009). Third, a shift needs to happen from the focus of evaluations measuring technical or managerial tasks to supporting and encouraging principals to face the adaptive ones (Heifetz &

Linsky, 2002; Micheaux & Parvin, 2018). All district principals and principal supervisors need to agree and take ownership of this shift; otherwise, it will not succeed (Tran & Bon, 2015). Some research indicates the importance of using self-reflective measures and a more formative approach to allow for principals to experience growth from the challenages that they are facing with the support of a principal supervisor (Close et al., 2018; Fuller et al., 2015; Welner & Burris, 2011). Fourth, principal evaluations should involve input from not only supervisors but also teachers, students, and parents (Liu et al., 2014). Incorporating these aspects would create a more accurate and authentic principal evaluation system.

How One Method Combines Current Theories of Principal Leadership and Current Policies

Researchers on principal leadership and principal evaluation methods agree that leadership behaviors are specific to the context and shared across the organization (Halverson & Kelly, 2017; Leithwood, 1994; Leithwood & Jantzi, 2008; Marks & Printy, 2003; Printy et al., 2009; Spillane, 2006; Spillane et al., 2004). First, some specific leadership behaviors have positively affected school culture, such as shaping the mission or setting the direction of the school (Hallinger, 2003, 2005; Leithwood et al., 1998; Leithwood & Sun, 2012; Marks & Printy, 2003; Robinson, 2008). Second, those behaviors are distributed across the organization, and school members work collaboratively towards their attainment (Halverson & Kelly, 2017; Leithwood & Jantzi, 2008; Marks & Printy, 2003; Printy et al., 2009; Spillane, 2006; Spillane et al., 2004). Third, the type of principal leadership behaviors exhibited may change over time (Day et

al., 2016; Marks & Printy, 2003). Finally, the context matters when examining the leadership behaviors needed and displayed (Day et al., 2016; Hallinger, 2005, 2018; Halverson & Kelly, 2017; Leithwood, 1994; Southworth, 2002; Spillane, 2006; Spillane et al., 2004; Urick & Bowers, 2014). Principal leadership is both socially and situationally distributed across the organization (Halverson & Kelly, 2017; Leithwood, 1994; Leithwood & Jantzi, 2008; Marks & Printy, 2003; Printy et al., 2009; Spillane, 2006; Spillane et al., 2004).

Methods of evaluation that are more accurate measurements of principal leadership have mirrored some of these agreed-upon qualities. First, researchers advocate for reliable and straightforward principal evaluation tools that are explicitly aligned to professional learning standards which provide clarity and leadership expectations (Corcoran et al., 2013; Fenton, 2016; Micheaux & Parvin, 2018; Sinnema & Robinson, 2012; Tran & Bon, 2015; Williams, 2015). Second, they should be aligned to district strategic goals and focus on instructional leadership activities (Corcoran et al., 2013; Leithwood et al., 2004; Sun & Youngs, 2009). Third, since leadership changes over time, an evaluation should focus on how principals adapt their leadership to the current context and challenges they face (Day et al., 2016; Micheaux & Parvin, 2018). Fourth, because of the socially distributed nature of leadership, principal evaluations should involve input from not only supervisors but also teachers, students, and parents (Liu et al., 2014). Principal evaluation systems should focus on the context of the administrator.

While many current principal evaluation systems fail to acknowledge many of those qualities, there is one principal evaluation method that is built on these principles, the Comprehensive Assessment of Leadership for Learning (CALL). The CALL is a

formative assessment that is designed to improve teaching and learning through measuring distributed leadership for learning practices (Mark H. Blitz & Modeste, 2015; Halverson et al., 2014). Written through a grant funded by the Institute of Educational Sciences in the US Department of Education, it draws on significant educational leadership studies to measure leadership practices in five domains. Principals participating in this research will take the survey as a self-reflection tool to assess their competency in these five domains. Since self-reflection is an essential aspect of adult learning, the CALL can serve as a formative assessment for school leaders (Boyatzis, Boyatzis, & McKee, 2005; Drago-Severson, 2009; Kegan & Lahey, 2016). Teachers and staff also have the opportunity to complete the survey, measuring the building administrators on the same subdomains (Blitz & Modeste, 2015; Halverson & Kelly, 2017). The multi-rater aspect of the CALL allows schools to have a complete picture of how leadership is exercised and perceived at the school, given its unique context (Halverson & Kelly, 2017). By providing opportunities to identify strengths and weaknesses in the school practice, leaders can make decisions about pathways for school improvement (Halverson & Kelly, 2017). It focuses on leadership at the macro and micro level by examining specific tasks and behaviors that capture levels of leadership from the perspectives of not only the school leaders but their staff as well (C Kelley et al., 2012)

The CALL survey addresses instructional leadership by assessing leadership tasks and providing formative feedback in three ways. First, it does so in the transparency of the design, as learning can occur while leaders are taking the assessment. Second, the assessment results are provided in a way that establishes levels of expertise, patterns in responses, and statistical calculations. Lastly, it gives leveled guidance on the next steps

to build distributed leadership for learning (Halverson & Kelly, 2017; Kelly et al., 2012). This formative feedback allows for leaders to grow as a result of taking the survey.

CALL Survey Research

The CALL draws its research from several key principles. Foundational to the CALL is the idea of leadership for learning, which is a theory of leadership that combines the leadership skills from instructional leadership, transformational leadership, and shared instructional leadership that has been proven to have the largest effect on student achievement (Hallinger, 2011; Murphy, Elliott, Goldring, & Porter, 2006). Leadership for learning emphasizes that leadership is enacted within a particular context, moderated by the values and beliefs of the leader, achieved through the influence of others, and directed towards student academic growth (Hallinger, 2011; Robinson, 2008). This research-based framework suggests that a focus on distributed leadership is a crucial way to drive school improvement (Halverson et al., 2014). The CALL synthesizes leadership practices from the research of four educational leadership studies:

(a) the seminal work of Hallinger and Heck (1998),

(b) the Vanderbilt Assessment of Leadership in Education (VAL-Ed),

(c) Marzano's *School Leadership That Works* (2005), and

(d) Bryk, Sebring, Allensworth, Luppescu, and Easton's (2010) work through the Consortium for Chicago School Research.

Hallinger and Heck (1998), in their seminal work, reviewed research from 1980-1995 that explores the relationship between principal leadership and student achievement. In doing so, they found that though indirect, principals exercise a small, statistically significant effect on school effectiveness and student achievement. Building on this work,

the Wallace Foundation (2001) reviewed ten years of research (2000-2010) on principal leadership through more than 70 reports that they published. In doing so, they identified five specific key responsibilities that effective principals demonstrate to increase student achievement. They are as follows:

(a) shaping a vision of academic success for all students based on high standards,

(b) creating a climate hospitable to education in order that safety, a cooperative spirit and other foundations of fruitful interaction prevail,

(c) cultivating leadership in others so that teachers and other adults assume their part in realizing the school vision,

(d) improving instruction to enable teachers to teach at their best and students to learn at their utmost, and

(e) managing people, data and processes to foster school improvement.

They note that these five key responsibilities must be done interdependently to achieve success. These two studies become the basis of research for determining what measures could be used to evaluate principals, such as is done in the VAL-ED.

The VAL-Ed is a survey of principal leadership that was created in 2008 through the Wallace Foundation to measure core components and key processes of principal leadership (Porter et al., 2008). It was one of the first of its kind to assess a principal's learning-centered leadership *behaviors* that have a direct impact on teachers' performance and student achievement. Leadership behaviors were explicitly evaluated as opposed to previous theories of leadership that focused on less objective methods such as leadership qualities or personality traits. Aligned to leadership standards, it follows a 360-

degree model where, in addition to the leader taking the survey, staff members from across the school complete it as well. Questions on the survey focus on six key standards:

(a) high standards for student learning,

(b) rigorous curriculum,

(c) quality instruction,

(d) a culture of learning and professional behavior,

(e) connections to external communities, and

(f) performance accountabilities.

It measures each of these standards through six key leadership processes: planning, implementing, supporting, advocating, communicating, and monitoring. The survey went through a validation process that included a nine-school pilot. After the pilot, interviews were conducted. Changes were made to this first version of the survey to make the scale numerical (1-5) and clarify the wording of the questions. The survey was then validated again nationally in 300 schools, and the second pilot produced more valid results (Porter et al., 2008). Psychometrics also were conducted that showed the VAL-ED to work well in a variety of settings, as it was found to be unbiased, reliable, feasible for widespread use. Additionally, the VAL-ED has a valid construct, provides accurate and useful reporting of results, yields a diagnostic profile for formative purposes, and can be used to measure progress over time (Goff, Salisbury, & Blitz, 2015; Porter et al., 2008).

The third study that was influential in the development of the CALL was Marzano's (2005) study, *School Leadership that Works*. Marzano synthesized a variety of studies on principal leadership to find twenty-one key features of a principal's practice that are correlated with student outcomes. Recently, Marzano's work has been called into

question due to the failure of a randomized controlled experiment to produce an impact on student achievement (Jacob et al., 2015). However, it should be noted that the sample size for the study was small (41 schools) and confined to one particular area of the country (Northern Michigan). Additionally, the study describes the delivery of professional development but not how that was acted upon in the schools. Thus, the twenty-one features could be effective when implemented appropriately. It is important to note that the CALL only draws from eight of the features. They are as follows:

(a) change agent, which describes how the leader challenges the status quo;

(b) cultures of achievement, the leader creates and promotes to guide the work of teaching and learning;

(c) discipline, which describes the responsibility for decreasing distraction during instructional time;

(d) input, which measures teachers' influence in running the school and building a shared sense of purpose and consensus;

(e) involvement in curriculum, instruction, and assessment practices;

(f) monitoring and evaluating, which is important in providing feedback to teachers and students;

(g) order in routinizing practices so that more effort can be spent on improving student learning; and

(h) resources that leaders distribute to shape desired practices.

While these eight behaviors are from Marzano's work, they are characteristics that are validated in other research studies, lessening the validity threat from the randomized controlled experiment (Kotter & Rathgeber, 2006; V. Robinson, 2011; Sebastian et al.,

2017; Urick, 2016). These eight factors influenced the CALL domains and subdomains (Halverson & Kelly, 2017).

Lastly, an additional piece of literature on which the CALL is based is the *Organizing Schools for Improvement: Lessons from Chicago study by the Consortium for Chicago School Research*. Drawing from the extensive longitudinal database about Chicago and its public schools, Byrk, Sebring, Allensworth, Luppescu, and Easton (2012) looked at student achievement data in reading and math based on the Iowa Test of Basic Skills; administrative data from the Chicago Public Schools Comprehensive Student Information System; and principal, teacher, and student surveys. These were analyzed how the organizations of schools and local communities affect their ability to improve student learning and engagement. The reform efforts at the time the was data collected allowed them to study this issue. They found that there were five essential supports for improving student learning. They are as follows:

(a) school leadership,

(b) parent and community ties,

(c) professional capacity of the faculty,

(d) school learning climate, and

(e) instructional guidance.

The five essential supports from the longitudinal work in Chicago were included in the creation of the CALL subdomains.

All four of these works were synthesized to identify the key leadership behaviors that link school leadership and student outcomes (Halverson & Kelly, 2017). Since many of the studies described practices, CALL worked to reframe these characteristics to

demonstrate how these practices distributed across the school context. The survey developers worked to make these broader themes more detailed through the development of five domains of school leadership practice. These five domains are as follows:

(a) focus on learning,

(b) monitoring teaching and learning,

(c) building professional community,

(d) acquiring and allocating resources, and

(e) establishing a safe and effective learning environment (Halverson & Kelly, 2017).

Under these five domains, the characteristics were expanded to create subdomains. The CALL subdomains synthesize four studies on effective leadership practices.

CALL Survey Validation Process

After the creation of the initial drafts of the CALL survey, an extensive validation process occurred which found the survey to be a valid method of assessment and formative feedback for principals (M. Blitz, 2012; Mark H Blitz & Salisbury, 2012; Camburn & Salisbury, 2012; Halverson & Kelley, 2017; Halverson et al., 2014; C Kelley et al., 2012; V. Robinson, 2011; Sebastian et al., 2017; Urick, 2016). Validations for the first working draft of the CALL were done in several ways (Camburn & Salisbury, 2012; Kelly et al., 2012). First, the instrument was piloted in over seventy schools in 2011 (Camburn & Salisbury, 2012). Two practitioner focus groups were created with individuals from different schools and in different positions (Kelly et al., 2012). These focus groups were tasked with reviewing and improving the draft survey design (Camburn & Salisbury, 2012). Then the North Central Regional Educational Laboratory

(NCREL) met with seventy-eight school-level leaders from eleven different middle school and high schools to conduct an item selection and construct validity inquiry (Kelly et al., 2012). In the second year, 2012, the survey was pilot tested with 1,784 educators in eleven districts in the Midwestern and Southern United States. Three rounds of interviews in six schools were conducted around school context, survey administration, and utility of feedback with the principals and other survey users, such as leadership team members, teachers, and other staff. The third round of the interviews included an interview with the building principal to discuss survey results (Kelly et al., 2012). Multiple stakeholders were involved in the CALL validation process.

Through these interviews, they discovered five aspects that were barriers related to the usability of the survey, some of which would later be changed (Blitz & Salisbury, 2012). First, the language needed to be more accessible; the terms needed to be ones that were actually used in practice in schools. Second, leadership that happened beyond school walls, such as district-level practices, needed to be addressed. Third, they found that there was a socially desirable response and that all participants of the survey worked to try to get the right answer instead of answering honestly. Fourth, in working with a 360-degree model, some items did apply to certain personnel more than others depending on their position. Lastly, identifying the appropriate practices and scales for items was a challenge as this survey aimed to measure behaviors and not personality traits. In response to this feedback, the CALL creators changed the language to focus more on practice and defined key terms to allow for differences in language. They also began developing a survey for central office supervisors to complete the 360-degree feedback

(Blitz & Salisbury, 2012). Changes were made to the CALL survey due to stakeholder feedback from the validation process.

The data from the second year of the pilot was used to begin the analysis of the statistical properties of the survey to inform further survey refinement through contracting with the University of Wisconsin Survey Center (UWSC) (Camburm & Salisbury, 2012; Kelly et al., 2012). Two-level hierarchical linear models were conducted, and all but three subdomain scores had a validity above .7 (Camburm & Salisbury, 2012). The initial reliability analyses had mixed results with Cronbach's Alpha reliability scores at .7 or above for only eleven of the twenty subdomains. For each scale with reliability below .7, those items were reviewed, and language was added or revised to improve reliability. Reliability analysis was then conducted on the Rasch scale, which was much more successful. A final step was to correlate the survey with the Interstate School Leaders Licensure Consortium Standards (ISLLC). Though the CALL is correlated because this was not a specific aim of the survey, the correlation is not one to one (Kelly et al., 2012).

Further CALL Research

After the initial validation research was conducted, follow up studies have used the CALL to measure and research principal leadership (M. Blitz, 2012; Mark H. Blitz & Modeste, 2015; Bowers et al., 2017; Goff et al., 2015; Carolyn Kelley & Dikkers, 2016; Min, Modeste, Salisbury, & Goff, 2016; Modeste, Hornskov, Bjerg, & Kelley, 2018). Some studies compared the CALL with other instruments that measure leadership (Goff et al., 2015). Other studies analyzed the difference in responses between the building leaders and teachers who took the study (Blitz & Modeste, 2015; Bowers et al., 2017;

Kelly & Dikkers, 2016; Modeste et al., 2018). Studies also looked to analyze how collaboration unfolds in a school building (Min et al., 2016). CALL research has affirmed the validity of the survey and expanded on its possible uses. This dissertation study builds on the research of two significant CALL studies mentioned below (See Figure 1 for a complete timeline).

Figure 1. Timeline of CALL Research Significant to this Dissertation Study.

One significant CALL study sought to determine whether teacher and principal responses to the same survey were correlated. In Blitz and Modeste's 2015 study, raw responses from 165 administrators and 3,663 teachers from 121 different schools were

standardized to make them z-scores. Twenty-one independent samples t-test were run to determine the differences between principal and teacher scores for each of the twenty-one domains. Overall, school leaders rated leadership practices higher than teachers except *supporting students who traditionally struggle* (1.4) (See Table 1). Blitz and Modeste (2015) ranked the twenty-one subdomains by quartiles, starting with the areas of agreed-upon effectiveness in quartile 1 and then ending with those with the largest differences in quartile 4. The quartiles allowed them to make comparisons across particular domains. The six domains in the 4th and 1st quartile are listed in Table 1 with the mean difference in parentheses. The means are positive where the teacher means are larger and negative, where the administrator means are larger. The positive and negative numbers are due to the standardization of the scores to zero. While domain differences did not follow a specific pattern, this data was not focused on school-level data, but all administrators and all teachers. Looking at individual pairings of teachers and administrators by school might have produced a more correlated response. Regardless, this data re-affirms the purpose of the 360-degree aspect of CALL that will allow leaders to discuss and develop ways to address discrepancies between perceived leadership.

Table 1

Mean Differences for Blitz and Modeste (2015) Study of Subdomains with the Largest and the Smallest Mean Differences Between Teachers and Administrators

Administrators Rated Themselves Significantly Higher than Teachers	**Teachers Rated Administrators Higher**	**Agreed Upon Effectiveness**
Sub 1.2: Formal leaders are recognized as instructional leaders (-.674)	Sub 1.4: Providing appropriate services for students who traditionally struggle (.163)	Sub 2.2: Summative evaluation of student learning (-.053)
Sub 3.3: Socially Distributed Leadership (-.591)		Sub 2.1: Formative evaluation of student learning (-.056)
Sub 5.2: Clean and safe learning environment (-.502)		Sub 3.2: Professional learning (-.215)
Sub 1.3: Collaborative school-wide focus on problems of teaching and learning (-.435)		Sub 5.3: Student support services provide a safe haven for students who traditionally struggle (-.217)
Sub 2.3: Formative evaluation of teaching (-.434)		Sub 3.4: Coaching and mentoring (-.225)
Sub 4.3: School resources are focused on student learning (-.419)		

A second study on CALL investigated the extent to which leaders and teachers agreed that these effective school leadership practices were occurring in their school. Using a congruency-typology model, Bowers, Blitz, Modeste, Salisbury, and Halverson (2017), ran a Latent Class Analysis on the 117 schools including 3,367 teachers and their school leader to find that there are three consistently different subgroups (low, moderate, high) among teachers and that these groups cluster non-randomly across the three types of schools (low leadership, moderate leadership, high leadership). For example, in high leadership schools, school leaders were the most experienced working with inexperienced teachers, but they had the greatest amount of agreement in scores. The non-random

clustering of this model suggests that there is more than one type of leader and that the leader may respond differently to the tasks based on the type of school. It also indicates that there is often a misalignment between how the leader views the leadership practices and how the teachers perceive them. This type of misalignment allows for intentional school improvement, particularly in areas of high disagreement. This dissertation study highlights the multifaceted nature of leadership.

Additional research is needed into the extent of difference in principal leadership as measured through responses on the CALL. First, many of the studies using CALL data have used the original set of CALL responses from the validation data (Blitz, 2012; Blitz & Modeste, 2015; Blitz & Salisbury, 2012; Bowers et al., 2017; Camburn & Salisbury, 2012; Min et al., 2016). While this data has a large sample size of teachers (n=3,767), the sample size of administrators is smaller (n =167) (Bowers et al., 2017; Min et al., 2016). Second, this data is not nationally representative, as it was from five states with most of the schools located in Wisconsin, Mississippi, and California (Bowers et al., 2017; Min et al., 2016). Third, the data is not multi-level because the data is not nested in schools. The unnested variables violate the assumption of independence as many staff will be rating one particular administrator and prevents them from distinguishing between types of teachers or leaders (Urick & Bowers, 2014). While other studies using CALL data have controlled for nesting variables, multi-level data was beyond the scope of this dissertation study (Bowers et al., 2017). Fourth, because the data is not aggregated at the school level, it has not yet taken into account context, such as what difference leading an urban school might make in a leader's practices and a staff's perception of their practices (Blitz & Modeste, 2015). Studies have also not looked at context as defined by the school level of

the school, such as secondary or elementary, as research on principal leadership indicates that the specific age group might make a difference in the leadership behaviors demonstrated (Hallinger, 2005, 2018). Finally, increased data is needed to see the difference in growth for student achievement outcomes between those using the CALL survey over traditional principal evaluation methods.

Context and the CALL

As this dissertation study looks at two specific areas of context through the barometer of the CALL survey, it important to note how the CALL addresses context. First, the CALL builds on research that was conducted specifically in an urban context, as the Five Essential Supports were developed at the Consortium for Chicago School Research that was conducted in Chicago Public Schools (Bryk et al., 2010; Halverson & Kelley, 2017). While the other three bases for research (Hallinger and Heck, Marzano, VAL-ED) were conducted in a mixed geographical context, the Chicago study found behaviors that were specific to urban leaders (Hallinger & Heck, 1998; Marzano, Waters, & McNulty, 2001; Porter et al., 2008). As this study looks at the context of region (rural, small town, suburban, urban), this is significant. Second, the CALL was developed considering the political shift that started with No Child Left Behind (NCLB) and currently continues with the Every Student Succeeds Act (ESSA) (Halverson & Kelly, 2017). Namely, Halverson and Kelly (2017) note that the expectations for principals have changed with them being more directly responsible for teaching practice and helping teachers and learners improve. They believe that this shift in expectations leads principals to be more distributed in their leadership in this current policy context than principals were in previous eras. This shift is noted in the research and development behind many of

the subdomain questions in the CALL (Halverson & Kelly, 2017). In fact, the survey questions themselves provide multiple choice responses on a Likert scale that increase in distributed leadership of the school's teachers and staff. The research based for the CALL incorporates context by containing research from an urban context (Chicago) and from the political context of leadership after NCLB.

The CALL survey specifically addresses context through questions on particular subdomains in Domain 4: Acquiring and Allocating Resources. The first subdomain, *personnel practices* (4.1), addresses the formal program in place to induct new teachers in the school. New teachers have high turnover rates, particular in urban contexts (R. Ingersoll & May, 2016; R. M. Ingersoll, 2001; R. M. Ingersoll & Smith, 2003). The CALL addresses this difference by asking specific questions about how personnel practices guide new teachers in integrating them into the school community. In *school resources focus on student learning* (4.3), the CALL asks specific questions around how extended learning opportunities during and after school day support struggling students. The rationale behind this set of questions was in research that showed inequities between urban schools and other contexts. In high performing districts, schools targeted resources toward high performers in extracurricular STEM activities, whereas in low performing, high poverty schools, students were more likely to get ineffective remedial learning as part of their extracurricular activities (Halverson & Kelley, 2017; V. Robinson, 2011). These questions are designed to examine leadership around student learning to ensure that they connect to instruction in meaningful ways. The last subdomain in Domain 4, *coordinating and supervising relations with families and external communities* (4.5), explores the school's relationship with the community itself. This subdomain addresses

how families are involved in supporting student learning. The research behind this question addresses the differences between middle class families and those families that are economically disadvantaged, as practices need to be adaptive to the unique needs of those contexts (Halverson & Kelley, 2017; Khalifa et al., 2016). These three subdomains incorporate research on the difference in these leadership practices due to context, particularly that of an urban context, to help leaders move to use more research-based practices.

Despite the research base behind the CALL and the three domains the specifically speak to the difference between urban schools and schools in other locations, a limitation of the survey itself is that it does not deal with specific contexts in every domain. While political context and regional context is mentioned in the research base that undergirds the CALL survey, context is not a focus of the survey. Instead, the CALL is focused on those practices which provide a direct link between leadership for learning and student achievement (Halverson & Kelly, 2017). The CALL only has one of the twenty subdomains that deals specifically with the community, a key factor of any context (Horsford, Grosland, & Gunn, 2011). This may effect some of the findings and interpretation of those findings. However, as the CALL is a research-based survey of twenty specific practices, it provides a constant measure to compare differences of how leaders prioritize leadership when determining the difference in means between the contexts of school level and region.

Theoretical Framework

Principal leadership is essential to measure due to the impact that principals have on student achievement and school climate (Brown & Wynn, 2009; Day et al., 2016; Greenlee & Brown, Jr., 2009; Johnson et al., 2012; Leithwood et al., 2004). While there is much emphasis on principal evaluation systems in the past few years, the systems are very general and do not provide opportunities for principal growth based on context (Halverson & Kelly, 2017). I chose the CALL for this dissertation study due to its unique framework, which measures specific leadership skills in a distributed fashion (Halverson et al., 2014). The measurement of these skills is so important because while we know that particular leadership behaviors are important, the uniqueness of each school context will dictate how principals prioritized those practices (Day et al., 2016; Hallinger, 2005; Halverson & Kelly, 2017; Southworth, 2002; Urick & Bowers, 2014). The individual multiple-choice items highlight the degree to which leaders and staff are involved in each of the tasks through their structured responses, illustrating the socially distributed nature of the leadership of the principal (Halverson & Kelly, 2017; Leithwood, 1994; Spillane, 2006; Spillane et al., 2004). Lastly, the CALL builds on the necessary components for adult learning. Adults grow best when allowed to reflect on where they need to improve (Boyatzis et al., 2005; Drago-Severson, 2009; Kegan & Lahey, 2016). Its formative nature allows for principals and their staff to reflect on specific leadership behaviors while providing concrete examples of how to improve those practices (Halverson & Kelly, 2017).

The study design allows me to examine how context affects leadership behaviors. I will be able to address some of the limitations of previous studies by using CALL survey data from 281 schools that were collected between 2016 and 2019. This sample is distinct from the one used in previous samples and will provide a comparison for how consistent the data from one sample set of teachers and leaders might be to another (M. Blitz, 2012; Mark H. Blitz & Modeste, 2015; Bowers et al., 2017; Goff et al., 2015; Carolyn Kelley & Dikkers, 2016; Min et al., 2016). In other words, I will be looking at how contingent leadership is by looking at data from different contexts (See Figure 2).

Basic Question:

To what extent do principal leadership practices vary based on context?

Research Proposal:

Difference between teacher and school leader perspectives in the 2016-2019 CALL dataset

Difference between elementary and secondary school teacher and leader perspectives

Difference between urban, suburban, and rural school leader perspectives

---

Hypothesized Results:

Comparison with CALL Blitz and Modeste (2015) data between domains will show different domains of agreed-upon effectiveness and different domains of larges differences

Elementary & secondary leaders perspectives' will not show a statistical difference in means. Elementary teachers and leaders perspectives will be less different than those of secondary teachers and leaders

Urban, suburban, and rural leaders will have statistically different subdomain averages

---

Implications:

Aid in our understanding of how teachers and administrators perceive skills and tasks associated with effective school leadership and whether or not their perceptions of those tasks align	Provide further evidence at the failure of current policies around evaluation systems to provide accurate measurements of the complexity of leadership	Allow district leaders and principals to examine the distribution of these subdomains across a school similar to their own, which would provide background for planning for school improvement	Provide suggestions for principal training programs and professional development through targeted support and resources

Figure 2. General Flow of Dissertation Framework.

I will be using the CALL data to analyze three different contexts: the 2016-2019 CALL dataset vs. the Blitz and Modeste (2015) dataset, school level, and region. In using the CALL to measure specific aspects of leadership, I look to examine which leadership for learning domains yields the greatest difference between teacher and school leader perspectives from the 2016-2019 CALL dataset. I will then compare the results of the 2016-2019 CALL dataset to the validation sample data used in Blitz & Modeste's 2015 study. This comparison will examine whether or not the domains of agreed-upon effectiveness between the leader and school staff are the same or different than those in the Blitz and Modeste (2015) sample (Blitz & Modeste, 2015). To make such a comparison valid, I will be replicating and extending that study by using a similar research question and following similar methods. This will allow me to examine those domains, which yield large differences to see how they compare to the Blitz and Modeste (2015) sample (Blitz & Modeste, 2015). The comparison is significant because the validation schools were in different states and the schools in a slightly different time than this sample. Consistency or lack thereof across the two samples would add to our knowledge of how context affects leadership.

Next, I will examine how context, as defined by the school level, effects teacher and administrator subdomain means. Since the size of the school and the level of pedagogy were found to be aspects that mitigate the effect size of leadership, it is important to measure this aspect of context (Hallinger, 2005; Leithwood & Jantzi, 2008; Nguni et al., 2006; Thoonen et al., 2011). The school level will be measured by examining how the difference in means between elementary and secondary teachers and administrators.

Third, I will examine context by disaggregating the data to determine how the region of the school affects responses on leadership subdomains. Research indicates a difference in principal leadership and principal evaluation systems due especially to the context of urban schools (Fuller et al., 2015; Greenlee & Brown Jr., 2009; Horsford et al., 2018; Hughes, Matt, & O`Reilly, 2015; Hunter & Donahoo, 2003; R. Ingersoll & May, 2016; Johnson et al., 2012; White et al., 2012). I will first classify school districts through their self-identified region of rural, small town, suburban, or urban. An ANOVA will be used to determine to what extent region creates differences between subdomain means.

This research will have significant implications for research, policy, and practice. Previous research tied principal leadership to specific theories and models of leadership, such as instructional, transformation, and shared instructional leadership (Bowers et al., 2017; Derrington, 2018; Hallinger, 2005, 2011, 2018b; Philip Hallinger & J Murphy, 1985; Leithwood & Riehl, 2005; Marks & Printy, 2003; Prestine & Nelson, 2005; Printy, 2008; Urick & Bowers, 2014b). However, those styles have significant overlap and do not fully encapsulate the different aspects of principal leadership (Blase, 1987; City et al., 2009; Day et al., 2016; Greenlee & Brown, 2009; Johnson et al., 2012; Leithwood et al., 2004; Leithwood & Riel, 2005). A few recent studies have examined leadership types to determine what these models look like in practice (Urick & Bowers, 2014; Bowers et al., 2017). This dissertation study aims to build on existing CALL studies to examine how principals choose to exercise leadership in different contexts as opposed to following a particular leadership style.

The idea of how principals' perceptions of their leadership match with teachers' perceptions of their leadership are significant to research as this data is self-reported.

While there has been much research on the indirect effects of principal leadership as measured through their actions that influence school and classroom conditions and teacher-perceived student outcomes, little research has been done on principal's perceptions of their own leadership (Hallinger, 2003, 2005; Hallinger & Heck, 1996; Leithwood, 1994; Leithwood & Jantzi, 2005; Southworth, 2002). As Urick and Bowers (2014) note,

> While much of the research has focused on teacher perceptions of leadership, little research has been done to examine principal perceptions of their own leadership practice and how those practices, in combination with school context, help to develop school conditions that support the inclusion of teachers as instructional leaders (Evans, 2010; Leithwood & Jantzi, 2008; Urick & Bowers, 2011).

Examining how principals perceive their own leadership allows us to examine their leadership style to see what aspects of leadership are prioritized due to context. Given the demographics of the comparison (location, school level, region), I will be able to examine the different types of principals to determine to what extent school level (elementary or secondary) and region influence their perspectives of leadership. It also allows for comparisons between teacher and principal perceptions. If the purpose of leadership is to make an impact on its followers, then how leaders and teachers perceive the same domains or skills is essential to examine the effectiveness of leadership in those particular areas (Bowers et al., 2017; Boyce & Bowers, 2018a; Reich, 1988; Urick & Bowers, 2019). Specific leadership practices, such as resource allocation and teacher evaluation, will be examined from the perception of the leader and the follower (Blitz & Modeste, 2015). The differences and similarities between these groups aid in our understanding of what skills and tasks are associated with effective school leadership, how teachers and administrators perceive those tasks, and whether or not their perceptions of those tasks

align (Bowers et al., 2017). This dissertation study aims to aid in our understanding of the interaction between teachers and principals around leadership practices and to what extent it is affected by context.

This research is unique in that it uses a recent data set. Previous studies of this nature were limited by the validation sample (Blitz & Modeste, 2015; Bowers et al., 2017). While not a random sample, this sample was collected to study leadership and make comparisons. Examining how context affects types of leaders builds on and extends previous leadership with a different data set (Blitz & Modeste, 2015). The different dataset is significant because if a context is a mitigating factor, the difference in means should be seen across multiple data sets using the same survey instrument.

This research also has significant implications for policy. While recent research has continued to identify leadership as multifaceted, and malleable, principal evaluation systems have remained static (Close et al., 2018; Corcoran et al., 2013; Derrington, 2018; Fuller et al., 2015; Piro et al., 2011; Sun & Youngs, 2009). Current models of principal evaluation are often unaligned with unique district characteristics and use general, flawed statistical measures to evaluate leadership (Corcoran et al., 2013; Shoaf et al., 2013; Sinnema & Robinson, 2012; Tran & Bon, 2015; Williams, 2015). The CALL provides an alternative to this flawed system (Blitz & Modeste, 2015; Halverson et al., 2014). The unique structure of CALL around the five leadership domains allows for specific leadership behaviors to be studied and compared. A difference in domain means due to the context will provide further evidence of the failure of current principal evaluation systems to provide accurate measurements of the complexity of leadership. The present study may be useful to state departments as they continue to revise principal evaluation

systems and growth metrics under the Every Student Succeeds Act (Close et al., 2018; Corcoran et al., 2013; Derrington, 2018; Fuller et al., 2015; Henig, 2013; Piro et al., 2011; Sun & Youngs, 2009). The present study will provide evidence of the need for more differentiated systems of principal evaluation and support. In the current system, where most principals are rated effective, it is more challenging to organize support, as means are not different enough to distinguish need (Herrmann & Ross, 2016; White et al., 2012). Greater differentiation across leadership domains allows for more significant comparisons and support for growth, which may include pairing with those that score higher across a particular domain with those that score lower (Ni et al., 2016). This research attempts to provide state departments with information as to the need to provide more detailed models of principal evaluation to incorporate the differences in leadership due to context.

Lastly, this dissertation study has many implications for practice as it provides both teacher and administrator rates on the same leadership for learning subdomains. Since the CALL is formative and focuses on leadership for learning skills and practices, data is useful for school improvement (Halverson & Kelly, 2017). By scrutinizing both teacher and leader responses, one can examine both the alignment and misalignment of teacher and leader practices (Blitz & Modeste, 2015). These conclusions will then allow district leaders and principals to explore the distribution of these subdomains across a school similar to their own, which would provide background for planning for school improvement. This data, disaggregated by school level (elementary or secondary) and region (urban, suburban, small town, rural), will also provide suggestions for principal training programs, professional development, and teacher and principal evaluation

systems (Bowers et al., 2017). In looking at which domains show the greatest differences for leaders between contexts, specific groups of principals can receive targeted support and resources. While there is some research around this, more is needed as policy and practice continue to treat principal leadership as monolithic and not dynamic and influenced by context (Bowers et al., 2017).

Research Questions

The purpose of this dissertation study is to assess the contingency of leadership based on context. I will examine and measure specific principal leadership practices as measured through the CALL subdomains through the following three groups of research questions:

1. To what extent do teacher and administrator perspectives differ based on ratings within the CALL assessment?
2. To what extent do elementary and secondary school administrator perspectives differ based on ratings within the CALL assessment? To what extent do elementary teacher and administrator perspectives differ based on ratings within the CALL assessment? To what extent do secondary teacher and administrator perspectives differ based on ratings within the CALL assessment?
3. To what extent do administrator's ratings differ between rural, small town, suburban, and urban schools based on ratings within the CALL assessment?

Chapter III

METHODS

The previous section described the research base that surrounds principal leadership, principal evaluation methods, and the Comprehensive Assessment of Leadership for Learning (CALL). This chapter describes the methods used in this particular study. This dissertation is a quantitative study that investigates to what extent context affects the prioritization of principal leadership tasks. For this dissertation study, context will be defined location (Blitz and Modeste (2015) data versus the 2016-2019 CALL dataset), through school level (elementary or secondary), and level of urbanization (urban, suburban, small town, rural). The purpose of the research is listed, followed by the study's research questions and hypotheses, the data collection methods, data analysis, and the ethical precautions taken.

Purpose

The purpose of this dissertation study is to identify how differences in context, as defined through two different data sets, school level, and region, change how educational leaders prioritize leadership tasks through the analysis for data from the Comprehensive Assessment of Leadership for Learning (CALL) survey. I chose the CALL for this dissertation study due to its unique framework, which measures specific leadership skills in a distributed fashion (Halverson et al., 2014). The measurement of these skills is so important because while we know that a specific set of leadership practices are essential,

the uniqueness of each school context will dictate how principals prioritize those practices (Day et al., 2016; Hallinger, 2005; Halverson & Kelly, 2017; Southworth, 2002; Urick & Bowers, 2013). The study design allows me to measure both the *socially* distributed and *situationally* distributed nature of leadership, as demonstrated through the Leadership for Learning framework of the CALL.

Researchers seem to agree that there are two ways in which leaders influence teachers and students. First, the leader is *socially* distributed across the organization, meaning that the work is not done by the principal alone, but through his or her orchestration of different key actors who contribute collaboratively to the achievement of school goals (Halverson & Kelly, 2017; Leithwood & Jantzi, 2008; Marks & Printy, 2003; Printy et al., 2009; Spillane, 2006; Spillane et al., 2004). Second, the leader is *situationally* distributed across the organization, creating or gathering different resources according to the context of the school (Halverson & Kelly, 2017; Leithwood, 1994; Spillane, 2006; Spillane et al., 2004; Urick & Bowers, 2014). Specific practices can be measured to identify the impact of principals and to learn how they adapt to different contexts. However, more research is needed on this topic (Halverson & Kelly, 2017; Leithwood & Sun, 2012; Robinson et al., 2008).

In using the CALL to measure specific aspects of leadership, this dissertation study examines leadership across different contexts. In the first analysis, I examined leadership for learning domains to find those that yield the greatest difference between teacher and school leader perspectives from the 2016-2019 CALL dataset. I then compared the results of this analysis to the validation sample data used in Blitz & Modeste's 2015 study (Blitz & Modeste, 2015; Blitz & Salisbury, 2012; Kelly et al.,

2012). The mean differences between teachers and administrators in the validation sample were compared to those in the 2016-2019 CALL dataset to see whether domains of large differences were similar or different. Next, this dissertation study examines context, as defined by elementary or secondary school, to determine if there is a difference in subdomain means. Since the size of the school and the level of pedagogy are identified by research to be significant aspects that mitigate the effect size of leadership, it is important to measure context as related to elementary or secondary schools (Hallinger, 2005; Leithwood & Jantzi, 2008; Nguni et al., 2006; Thoonen et al., 2011). Finally, context will be examined by disaggregating the data by region to determine how the region of the school effects responses on leadership subdomains. Research indicates a difference in principal leadership and principal evaluation systems due especially to the context of urban schools (Fuller et al., 2015; Greenlee & Brown, 2009; Horsford et al., 2018; Hughes et al., 2015; Hunter & Donahoo, 2003; Ingersoll & May, 2016; Johnson et al., 2012; White et al., 2012). For this data sample, survey participants self-identified their school as rural, small town, suburban, or urban. Districts will be separated into these four groups to examine how their school leader scores differ from each other. These three analyses allow me to investigate how administrators prioritize different leadership practices due to context.

The purpose of this research is to assess the differences in specific principal leadership practices between different contexts. The main research question is to what extent do principal leadership practices vary based on context?

Other research questions are as follows:

1. To what extent do teacher and administrator perspectives differ based on ratings within the CALL assessment?
2. To what extent do elementary and secondary school administrator perspectives differ based on ratings within the CALL assessment? To what extent do elementary teacher and administrator perspectives differ based on ratings within the CALL assessment? To what extent do secondary teacher and administrator perspectives differ based on ratings within the CALL assessment?
3. To what extent do administrator's ratings differ between rural, small town, suburban, and urban schools based on ratings within the CALL assessment?

Sample

This dissertation study is a secondary analysis of the Comprehensive Assessment of Leadership for Learning data. Different districts and states have chosen to contract with the CALL to take this survey. As part of their participation in the CALL, they agree that the de-identified data may be used for future research. The researchers at the University of Wisconsin-Madison collected this data. I met with the researchers three times via Zoom in 2018 and 2019. The first time was to discuss the possibility of this study as well as ways to use the CALL survey to research principal leadership (M. Blitz, personal communication, August 14, 2018). The second time we met, we discussed the particulars of the present study. The CALL Survey Team at the University of Wisconsin-Madison generously allowed me to use the CALL data for this dissertation study (M. Blitz & R. Halverson, personal communication, December 3, 2019). The CALL Survey

Team at the University of Wisconsin-Madison downloaded the data from the CALL survey and sent me the raw files. We met a third time to discuss how to read the data and the differences between the files (M. Blitz, personal communication, August 14, 2018). Thank you to the CALL team for their generous donation of time and for allowing me to use the CALL survey to study principal leadership.

Due to the many different versions of the CALL survey, this dissertation study only uses data from the most current version of the CALL at the time of the present study, hereafter referred to as the 2016-2019 CALL dataset. Participants took this survey online between the years 2016 and 2019. During that period, some districts contracted with the CALL to create a new survey that shifts some of the domains. That data has not been included. Data has been de-identified, leaving only the school name, organization name, and state. In taking this survey, participants classified themselves as administrators, teachers, or support staff. I deleted support staff responses as they were outside the scope of this dissertation study. The survey was divided into two different data files, one of which was elementary and the other secondary, allowing for elementary and secondary means to be compared. These two data files were combined to create one data set. Questions that were unique to the elementary or the secondary data set were kept separate (See Appendix A & B). Region was self-identified by the survey participants as rural, small town, suburban, or urban. The data includes 281 schools with 536 administrators and 8,027 teachers.

Several descriptive statistics were run on the data to examine the depth of the sample. Administrators and teachers are from sixteen states with the largest samples from Wisconsin (2632), Georgia (1724), and Illinois (1307). The smallest samples are from

South Carolina (18), Michigan (11), and Washington (27) (See Table 2 and Figure 3). Interestingly, there are no administrators represented from Michigan and South Carolina, the two smallest state samples. In South Carolina, no survey participant who identified as an administrator completed the survey. In Michigan, one administrator completed the survey in the second administration of the survey; however, only responses from the first time a school administered the survey were included.

Table 2

Survey Participants by State as Represented in the Total Sample, Teacher Sample, and Administrator Sample

	Total Sample		Teachers		Administrators	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
Arizona	193	2.3	179	2.2	14	2.6
California	513	6.0	473	5.9	40	7.5
Georgia	1724	20.1	1626	20.3	98	18.3
Illinois	1307	15.3	1193	14.9	114	21.3
Indiana	367	4.3	313	3.9	54	10.1
Michigan	11	.1	11	.1	0	0
Minnesota	504	5.9	493	6.1	11	2.1
Nevada	153	1.8	147	1.8	6	1.1
New Jersey	96	1.1	95	1.2	1	.2
New Mexico	34	.4	30	.4	4	.7
New York	531	6.2	511	6.4	20	3.7
South Carolina	18	.2	18	.2	0	0
Texas	403	4.7	386	4.8	17	3.2
Virginia	50	.6	36	.4	14	2.6
Washington	27	.3	26	.3	1	.2
Wisconsin	2632	30.7	2490	31.0	142	26.5
Total	8563	100.0	8027	100.0	536	100.0

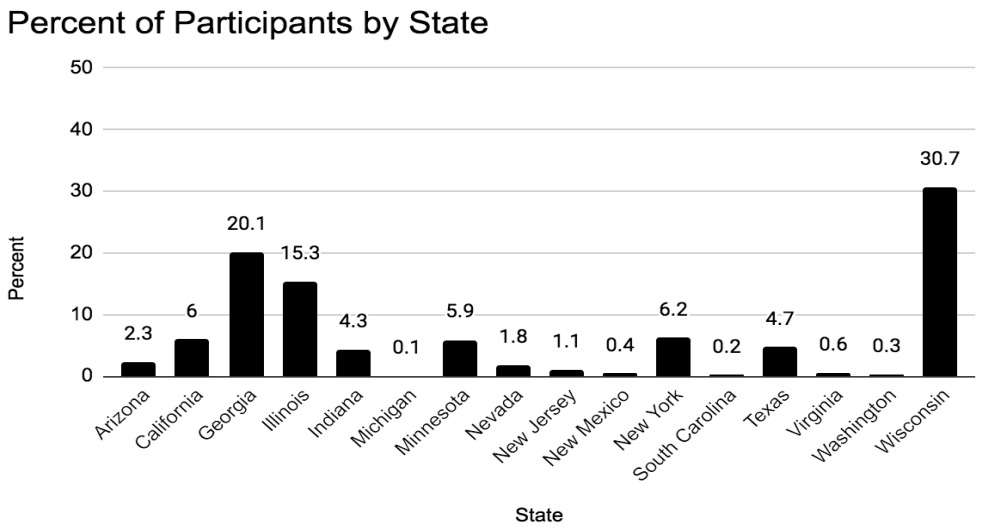


Figure 3. Percentage of Survey Participants by State.

The school level and geographical region also identify survey participants. There are 333 elementary school administrators and 203 secondary administrators (See Table 3). There are 4,206 elementary teachers and 3,821 secondary teachers. For those participants that identified their geographical region, 41.1% were from an urban area, 13.5% were from a suburban area, 8.3% were from a small town, and 16.8% were from a rural area (See Table 4). When looking at the number of schools in each location in this data sample, the urban schools are nearly twice as many as the next largest group, rural (n =58), with 109 schools (See Figure 4). Schools that identify as a suburban or small town school are roughly the same number with 30 and 26 schools, respectively.

Table 3

Number of Participants by Administrator and Teacher across Elementary and Middle or High Schools

Role Name	Total Sample		Elementary Schools		Middle or High Schools	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
Administrator	536	6.3	333	7.3	203	95.0
Teacher	8,027	93.7	4206	92.7	3821	5.0
Total	8,563	100.0	4539	100.0	4024	100.0

Table 4

Number of Participants by Administrators and Teachers across Regional Groups

Region	Frequency-Administrators	Frequency-Teachers	Frequency-Total	Percent
Rural	73	1,364	1,437	16.8
Small Town	28	684	712	8.3
Suburban	46	1,109	1,155	13.5
Urban	268	3,253	3,521	41.1
Total	415	6,410	6,825	79.7
Not Reported	121	1,617	1,738	20.3
Total	536	8,027	8,563	100.0

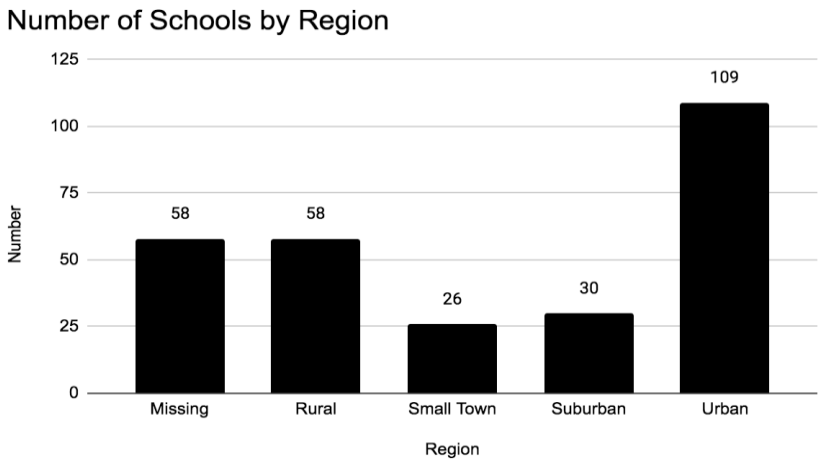


Figure 4. Number of Schools by Region.

Data Collection

The Comprehensive Assessment of Leadership for Learning (CALL) was chosen for this dissertation study because it measures specific leadership for learning practices (Halverson et al., 2014). Written in 2009 through a grant from the United States Department of Education, it focuses on leadership at the macro and micro level by examining specific tasks and behaviors that capture levels of leadership from the perspectives of not only the school leaders but their staff as well (Kelly et al., 2012). The CALL survey is intended to be given to three separate groups of stakeholders: the principal, the principal's supervisor, and the teachers that the principal supervises (Halverson & Kelly, 2017). This 360-degree format affirms current research and allows the principal to evaluate their leadership from different angles (Goldring et al., 2015; Liu et al., 2014; Shoaf et al., 2013; Tran & Bon, 2015). The CALL synthesizes leadership practices from the research of four key studies. First, the seminal work of Hallinger and Heck (1998) identifies five key responsibilities of leaders for improving schools. Second, the Vanderbilt Assessment of Leadership in Education (VAL-Ed), which highlights six core components and six key processes (Porter et al., 2008). Third, it incorporates Marzano's 21 key features of principal practices from his book, *School Leadership That Works* (2005). Lastly, the CALL includes the five essential supports from Bryk, Sebring, Allensworth, Luppescu, and Easton's (2010) work through the Consortium for Chicago School Research that discusses the critical organizational structures and practices for improved student success. The CALL incorporates research on principal leadership in a 360-degree survey that serves as a formative assessment.

The CALL went through an extensive validation process through which it was found to be an effective instrument to measure leadership (Blitz, 2012; Blitz & Salisbury, 2012; Camburn & Salisbury, 2012). Based on rubrics that were created by Halverson in conjunction with the University of Pittsburgh Institute for Learning (IFL), the CALL research team developed the initial draft of the survey (Kelly et al., 2012). The validation process included a seventy school pilot, two practitioner focus groups, an item selection and construct validity inquiry through the North Central Regional Educational Laboratory (NCREL) (Blitz, 2012; Blitz & Salisbury, 2012; Camburn & Salisbury, 2012; Halverson et al., 2014; Halverson & Kelly, 2017; Kelly et al., 2012). It was pilot tested with 1,784 educators in eleven districts in the Midwestern and Southern United States. The pilot was followed by three rounds of interviews, which led to the changing of language to focus more on practice, the definition of key terms, and the development of a survey for central office supervisors, which resulted in complete 360-degree feedback (Blitz & Salisbury, 2012). The University of Wisconsin Survey Center (UWSC) analyzed the statistical properties of the survey to inform further survey refinement through contracting with (Camburm & Salisbury, 2012; Kelly et al., 2012). UWSC conducted two-level hierarchical linear models, and all but three subdomain means had a validity above .7 (Camburm & Salisbury, 2012). While the initial reliability analyses had mixed results with Cronbach's Alpha reliability scores at .7 or above for only eleven of the twenty subdomains, reliability was increased for each scale with reliability below .7 by adding or revising language to improve reliability. Reliability analysis was then conducted on the Rasch scale, which was much more successful (Camburm & Salisbury,

2012). A final step was to correlate the survey with the Interstate School Leaders Licensure Consortium Standards (ISLLC). The CALL was found to be correlated with the ISLLC, but because this was not a specific aim of the survey, the correlation is not one to one (Kelly et al., 2012). The CALL validation process proves the CALL to be a reliable assessment.

The CALL has five domains and twenty subdomains, as seen in Table 5. These subdomains will be used as points of comparison between groups throughout the study. This chart gives the organization of the CALL survey and will serve as a framework for results throughout the study. When specific subdomains are mentioned in the results or the conclusion, the text of the subdomain will be listed in italics with the subdomain number in parentheses to increase readability. For example, Subdomain 1.4 will be written *providing appropriate services to students who traditionally struggle* (1.4).

Table 5

CALL Domains and Subdomains with the Number of Items in the Survey per Subdomain

CALL Domains and Subdomains					
Domains:	**1: Focus on Learning**	**2: Monitoring Teaching and Learning**	**3:Building Nested Learning Communities**	**4: Acquiring and Allocating Resources**	**5: Maintaining and Safe and Effective Learning Environment**
Subdomains:	1.1 Maintaining a school-wide focus on learning (7 items)	2.1 Formative evaluation of student learning (11 items)	3.1 Collaborative school-wide focus on problems of teaching and learning (8 items)	4.1 Personnel practices (6 items)	5.1 Clear, consistent and enforced expectations for student behavior (18 items)
	1.2 Formal leaders are recognized as instructional leaders (4 items)	2.2 Summative evaluation of student learning (7 items)	3.2 Professional learning (9 items)	4.2 Structuring and maintaining time (7 items)	5.2 Safe learning environment (11 items)
	1.3 Collaborative design of integrated learning plan (8 items)	2.3 Formative evaluation of teaching (11 items)	3.3 Socially distributed leadership (7 items)	4.3 School resources are focused on student learning (8 items)	5.3 Student support services provide a safe haven for students who traditionally struggle (12 items)
	1.4 Providing appropriate services for students who traditionally struggle (11 items)	2.4 Summative evaluation of teaching (13 items)	3.4 Coaching and mentoring (4 items)	4.4 Integrating external expertise into the school instructional program (3 items)	
				4.5 Coordinating and supervising relations with families and the external communities (10 items)	

Data Analysis

Before completing any analysis, I screened the CALL data. First, the CALL research group at the University of Wisconsin- Madison collected data. The researchers downloaded the data and sent it to me in raw form. Participants took this survey online between 2016 and 2019. During this period, some districts contracted with the CALL to create a new survey that shifts some of the domains. That data has not been included. The elementary and secondary files were separate with different questions in each file. Each file was sorted by the participant's role. I deleted those participants who identified as "support staff" as this role was outside the purpose of this dissertation study, which left only those participants who identified as "teacher" or "administrator." While the literature refers to school leaders as principals, the CALL survey does not differentiate between the principal or assistant principal, calling them all by the term "administrator." Because of this, when discussing the methods or results of this study, I will use the term "administrator." However, when referencing the literature, I will continue to use the term principal.

Many schools contracted with the CALL group to administer the survey once or twice a year over one to three years. Each file contained all of the times that the survey was given at a particular school. A timestamp identified when the participant started the survey and when they finished. As the survey is formative, each time that an individual takes the survey, their scores should rise slightly. Given that fact, I only included the first time each school group took the survey in this dissertation study. While most schools took the survey separated by many months, at times, it was difficult to determine where one survey administration ended, and another began in a particular school. I defined the

first administration of the survey as all of the participants who took the survey within a month of the first participant from their school. Any subsequent responses from that particular school were deleted, leaving the sample of 8,563 participants. The 2016-2019 CALL dataset contains only the data from the first time participants completed the CALL survey.

The two files, elementary and secondary, needed to be combined to analyze the data. First, I went through each file to create duplicate columns for questions that were not in one of the two databases. For example, for question q010422ES(!), a column was in the secondary data file under that name. I did the reverse when secondary questions differed from elementary questions. Some questions contain a school level code (ES for Elementary and HS for secondary), but some did not (For a full list of questions, see Appendix A). In Appendix A, I list each question by its variable number, the school level of the question (elementary, middle/high school, or both), the role (teacher, principal, or both), if it is a filter question, and whether it contains "I don't know" responses. Then both data sets were combined to create one data set.

After combining the data set, I removed all the filter questions. Some questions in the CALL survey examine whether or not a particular practice is present in a school before asking subsequent questions about that practice's effectiveness. While most of the CALL survey questions are on a scale of 1 to 5, these questions contain a "no" answer. If the survey participant answers "no" to a filter question, they are given a "-1." They are automatically given a "1" for subsequent related questions. On the recommendation of the CALL staff, I deleted all filter questions. Each of these deleted questions contained "-

1" no responses except for q030130A, where no administrator rated the practice as not present. I deleted this question as well (See Appendix A).

The survey data was then imported into SPSS software. Since the CALL is formative in its design, each question is built on an ordinal scale from 1 to 5 that gives a range of answers that indicate different practices which might be done by a leader to complete the task (Blitz, 2012; Blitz & Salisbury, 2012; Camburn & Salisbury, 2012; Halverson et al., 2014; Halverson & Kelly, 2017; Kelly et al., 2012). Scores were converted from a 1-5 scale to a 0-4 scale because the number zero has value. Some questions contained "I don't know" responses. Theses responses were often coded with a (!). The "I don't know" responses were given a score of zero in the data. When the scores were converted to a 0-4 scale, I coded those scores that were already at zero as missing in this conversion, as was recommended by the CALL team. As the survey was elective for participants, not all participants completed all of the questions. Where data was missing, the cell in the original data file was left blank. In the data transformation, SPSS continued to leave these cells blank. This data transformation resulted in incomplete data. However, there was still a large number of responses to each question.

Following the previous research using the CALL survey data, I averaged the survey items by subdomains to form a subdomain score for each survey participant to create the twenty CALL subdomains (Blitz & Modeste, 2015; Bowers et al., 2017). Questions are coded by subdomain with the first two digits of the question number being the domain number, the second two digits being the subdomain number, and the remaining digits the question number. For example, q020460 is subdomain 2.4, question 60. I averaged all of those questions, which had the same subdomain number to create a

subdomain average following the recommendations of the previous research using the CALL data (See Appendix B & C). In Appendix B, I list each of the questions included in each subdomain, separated by teachers and administrators. In Appendix C, I list descriptive statistics for each of the twenty subdomains, including the number of participants, the minimum score, the maximum score, the mean, and the standard deviation.

As the survey is formative, the range of answers for each question indicates the range of practices that characterize that task (Halverson & Kelly, 2017). For example in testing the leadership practices around English language learners (ELL) in a school, the survey might say "No one takes primary responsibility for teaching this students" under the lowest response (0) and "The English language learner teacher, the classroom teacher, and the support team work together to develop the learning plans" under the highest response (4). Thus, the survey items themselves provide formative feedback by illustrating what schools could do to improve on a particular task. On this Likert-style survey, zero is always the minimum as it is the lowest score, and four is always the maximum. However, the mean scores of each subdomain vary. For example, some subdomains, such as *formal leaders are recognized as instructional leaders* (1.2), have a high mean score of 2.947 for the overall sample of teachers and administrators. Other subdomains, such as *coaching and mentoring* (3.4), have a low mean score of 1.511. The scores themselves represent a range of the frequency with which teachers and administrators believe that leadership for learning practices are found in their schools.

These twenty subdomain means were converted to z-scores to standardized responses for both groups of respondents (See Appendix D). In Appendix D, I list the

number, minimum score, maximum score, mean, standard deviation, skewness, and kurtosis for each of the twenty subdomains. Standardized z-scores distribute along a z-distribution, which centers the scores to a mean of zero. For example, instead of distributing from 0 to 4, z-scores now distribute from a minimum score of -3.5 to a positive score of 2.5 with 0 as the mean and a consistent standard deviation of one. I used z-scores following the recommendations of Blitz and Modeste (2015) to allow me to compare subdomain means similarly to the previous study.

Z-scores are used to show the score's relative position or where a score is in relation to the rest of the scores in the distribution (Mertler & Reinhart, 2017). Z-scores are appropriate for this analysis as they allow for a more accurate comparison of mean scores. As Blitz and Modeste (2015) write,

> Standardized z-scores are values that have been standardized along the normal distribution, also known as the z-distribution. In CALL, this process of standardizing the respondents' scores along the normal distribution places all of the scores into dimensionless units above and below the mean of the data. Standardizing the variables allows us to compare subgroup differences across sub-domains. Leaders may be one standard deviation higher than teachers on a given sub-domain and one standard deviation lower than teachers on another sub-domain. We standardized the data to create a mechanism through which we can make comparisons, however limited, across sub-domain scores. (pp.357-368)

The z-scores allow for comparisons between groups across sub-domain means. Z-scores are listed by teacher and administrator separately in Appendix E. Appendix E displays the number, minimum score, maximum score, mean, standard deviation, skewness, and kurtosis for each of the twenty subdomains. The descriptive statistics are displayed first for teachers and then for administrators. Means no longer read zero as the sample was split. Administrator means are now consistently higher than teacher means.

This research will analyze the extent to which differences between groups exist with an effect measure. The results of each of the analyses are reported in effect sizes. Effect sizes are generally expressed as the standardized mean difference between an experimental and control group (Hendriks & Scheerens, 2013). While effect sizes will vary in significance depending on the type of statistical analysis run, in educational research generally an effect of 0.20 is considered small, 0.40 is considered moderate, and 0.60 is considered large (Cohen, 1992; Hattie, 2009). The effect size measurements will be reported for each analysis below, as ANOVA tests have a slightly different effect size measurement. For the t-test analyses, I use Cohen's d to measure effect sizes. Partial eta squared is used for the ANOVA. Effect sizes will show the extent to which the difference is significant between the two groups measured in the analysis.

Differences between the 2016-2019 CALL Dataset and

Blitz and Modeste (2015) Data

Three types of tests were run in SPSS Version 25 to test the hypotheses. The first major research question is to determine to what extent distributed-leadership domains differ between teacher and school leader perspectives based on ratings within the CALL assessment. I ran an independent samples t-test on the standardized z-scores. This first question seeks to replicate and extend the Blitz and Modeste (2015) data. I follow their recommendations for comparisons between the domains, including ranking subdomain differences in quartiles to make comparisons between the largest and smallest differences between teacher and administrator means (See Appendix F). In Appendix F, I list the mean difference, standard error difference, upper and lower confidence intervals of difference, and quartiles. Subdomains are listed in rank order from those that have the

largest difference in means between teachers and administrators to those that have the smallest difference in means. I ran twenty independent sample t-tests on administrators' and teachers' responses across each of the twenty standardized subdomain scores to analyze the significant difference between the two group means. The mean difference scores were computed in the output (Blitz & Modeste, 2015). A t-test is an effective measure of mean differences when the independent variable is defined as having two categories, in this case, principals and teachers, and the dependent variable is quantitative, in this case, the mean from each subdomain (Mertler & Reinhart, 2017; Schreiber & Asner-Self, 2011). The intent is to have at least 64 members in each independent group at $\alpha = .05$ to detect a moderate difference (Cohen, 1992). As this sample is over 8,000 participants, the t-test met this threshold. The effect size is measured as follows: .00-.20 (weak or no effect), .20-.40 (small effect), .40-.60 (moderate effect), and .60 or above (large effect) (Cohen, 1992). Theses means are then compared to the mean differences for each subdomain as recorded in Blitz & Modeste's 2015 study to determine if the domains of agreed effectiveness are the same in this sample as they were in the validation sample.

Differences between Elementary and Secondary Perspectives

For the second set of major research questions that deal with differences in leadership due to the school level, I ran an independent samples t-test to examine to what extent elementary and secondary school leadership perspectives differ based on ratings within the CALL assessment. Since respondents must identify whether or not they work at an elementary or a secondary school as one of the first questions on the CALL survey, this allowed for the groups to be separated (Halverson & Kelly, 2017). For the first

analysis in this question, the data for administrators was separated from the entire sample so that I could compare administrators' means. Then, I separated school leadership means into two groups (elementary, secondary) based on that demographic question. Using the z-scores for each subdomain, I compared means for each group of administrators through an independent samples t-test. As previously discussed, a t-test is an effective measure of significant mean differences when the independent variable is defined as having two categories, in this case, elementary principals and secondary principals, and the dependent variable is quantitative, in this case, the mean from each subdomain (Mertler & Reinhart, 2017; Schreiber & Asner-Self, 2011). The intent is to have at least 64 members in each independent group at $\alpha = .05$ to detect a moderate difference in this comparison (Cohen, 1992). As this sample is over 500 participants, the t-test met this threshold. The effect size is measured as follows: .00-.20 (weak or no effect), .20-.40 (small effect), .40-.60 (moderate effect), and .60 or above (large effect) (Cohen, 1992).

The second part of the school level analysis includes an independent samples t-test that examines secondary and elementary teacher and administrator responses to see whether teacher and principal subdomain means are more statistically significant between elementary or secondary schools. Returning to the full data set, I split the file into two groups, secondary and elementary. I ran an independent samples t-test. Since the file was split, the t-test examined how elementary teacher and administrator subdomain means differed and how secondary teacher and administrator subdomain means differed at the same time by running an independent samples t-test on each of the twenty subdomains for both groups. The effect size is measured as follows: .00-.20 (weak or no effect), .20-

.40 (small effect), .40-.60 (moderate effect), and .60 or above (large effect) (Cohen, 1992).

Differences Between Administrator Means Due to Region

For the third major research question, a one-way analysis of variance (ANOVA) is run on means for each of the districts as determined by their regional context. ANOVA tests are used to determine the variance in scores for more than two groups (Cohen et al., 2003; Mertler & Reinhart, 2016; Schreiber & Asner-Self, 2011). To examine that in this research, I separated school districts by their self-identified classification of urban, suburban, small town, and rural. Z -scores were analyzed through a one-way ANOVA to examine group mean differences. I ran an ANOVA on these mean averages with a p-value of $\alpha = <0.05$. An effect size of .10 will show a small effect, .25 a moderate size effect, and .40 a large effect (Cohen, 1992).

Familywise Error Rate

To control for a Type I error, the Familywise Error Rate (FWER) needed to be addressed. The familywise error rate is defined as the probability of coming to at least one false conclusion through multiple comparisons as done here. In any large sample with multiple tests, some will be statistically significant randomly due to the size of the sample. To control for this, I use the Bonferonni adjustment (Cohen et al., 2003). Because both of the first two analyses contain independent samples t-tests, I calculated this by hand. Since I am conducting 20 hypothesis tests, the level of significance cutoff is 0.05/20, which equals $p<0.0025$ (Cohen et al., 2003). In essence, to be a significant

difference, the p-value would have to be less than 0.0025. For the third analysis, I ran the Bonferonni adjustment through the post-hoc comparison t-test.

Human Participants and Ethics Precautions

Several ethical precautions are taken in this survey. When taking the CALL survey through the University of Wisconsin-Madison, participants agree that their responses may be used for research. In accordance with the guidelines of Columbia University, Teachers College, regarding the protection of human participants, a request for review was submitted and approved by the TC Institutional Review Board (See Appendix P).

Chapter IV

RESULTS

Introduction to the Results

This chapter is divided into three main sections that each discuss one of the three research questions of this dissertation study and the findings of the analysis in response to that question. The first section and analysis use an independent samples t-test that compares teacher means with administrator means for responses on each of the twenty subdomains of the Comprehensive Assessment of Leadership for Learning (CALL). I compare these means to the results of Blitz and Modeste's 2015 study. The second section includes an analysis that compares the subdomain means of elementary and secondary principals and teachers to test the extent to which there is a difference in the means between these two groups. The teacher and principal means are then compared between the elementary group and the secondary group, to see which of these two has less difference in their means and where that difference lies. I calculate Cohen's d to determine the effect size of the different groups in the analyses mentioned above. The third analysis compares subdomain means for administrators based on the contexts of rural, small town, suburban, or urban. In this analysis, I measure effect sizes using eta-squared.

Each of the comparisons and their effect sizes will be displayed on a heatmap, which in this study is referred to as a table with a heat gradient, that replicates the structure of Table 5. A heatmap is a data visualization tool that will allow the reader to

quickly see where the differences lie between subdomains (Bowers, 2010). As Bowers (2010) writes, "A heatmap takes tables of clustered numbers, which the human mind can not easily interpret for pattern recognition, and converts the table into blocks of color, aiding the human eye in visualizing patterns" (p. 6). I use gradients of the colors red and blue to show effect sizes between each of the comparisons. By showing the results for each analysis visually, the tables will provide clarity to the reader and aid in interpretation.

The purpose of this dissertation study is to examine how context effects ratings on the leadership for learning domains. Looking at which specific subdomains are different provide evidence on which leadership for learning behaviors principals prioritize in which contexts. Principal leadership is multifaceted and nuanced, making it challenging to tailor preparation programs, principal evaluation methods, and professional development to the needs of principals. Yet, the results of this dissertation study show specific domains that principals struggle to implement, which will allow for more targeted support.

The first research question extends the work of Blitz and Modeste (2015) by using similar methods to compare differences in principal leadership between teachers and administrators. The comparison is meaningful for two reasons. First, the results affirm Blitz and Modeste's finding that teacher and administrator responses on the CALL survey are different. A standard definition of leadership is influence over followers (Bowers et al., 2017; Boyce & Bowers, 2018a; Reich, 1988; Urick & Bowers, 2019). Therefore, to properly evaluate the quality of that principal leadership, it is necessary to have feedback from teachers. The need to have teacher perspectives is particularly true since the results

of the teachers are different than the principals'. The results of this analysis which show that teachers have different responses than principals indicate that current principal evaluation systems are missing a key piece of data they could use for improvement, which is including how teachers view their administrator's leadership (Bowers et al., 2017; Boyce & Bowers, 2018a; Reich, 1988; Urick & Bowers, 2019). Second, comparing the 2016-2019 CALL dataset to the Blitz and Modeste (2015) data confirms previous data that principal leadership does not follow a particular leadership style, rather different leadership practices are prioritized due to context. Systems designed for principal evaluation and growth should recognize the impact of context and factor those differences into leadership behavior prioritization (Halverson & Kelly, 2017; Leithwood & Jantzi, 2008; Marks & Printy, 2003; Printy et al., 2009). Recognizing the impact of context, the next two research questions examine how specific contexts (school level and region) affect how leadership tasks are prioritized. The results of these two analyses provide specific areas where systems of support for administrators, such as administrator preparation programs, professional development, and evaluations can be tailored to help principals grow in each of these specific contexts.

Comparison of the 2016-2019 CALL Dataset to Blitz and Modeste (2015) CALL Data

The first analysis addresses the question: **To what extent do teacher and administrator perspectives differ based on ratings within the CALL assessment?** The purpose of the analysis is to illustrate how leaders and teachers perceive leadership subdomains differently, considering their different contexts. None of the t-tests include

all of the sample participants due to the presence of missing data (See Appendix E). There are two reasons for the missing data. First, schools that contracted with the CALL administered the survey for school improvement. All survey participants were able to skip a question or exit from the survey before completing all of the questions. The skipped questions left data for some participants incomplete. Second, the CALL survey is written with some questions that contain "I don't know" or similarly worded responses. Following the recommendation of the CALL team, these responses were coded as missing data for this analysis (M. Blitz, personal communication, December 23, 2019). While the data could have been recoded, since the missing data was less than 5% for each subdomain in the teacher group, missing data were deleted analysis by analysis. Only half of the subdomains for administrators were below this 5% threshold; however, to keep the data consistent between the principal subdomains and the teacher subdomains, missing data were deleted analysis by analysis and not replaced or transformed (Strayhorn, 2010). Cases were excluded analysis by analysis, which causes the number of leader and teacher responses to vary from t-test to t-test.

Means were converted to z-scores to normalize the different distributions and to enable comparison between the two means. Z-scores create positive and negative numbers because they standardize a group of means to zero. When using z-scores, a negative or positive number shows the number's direction to the mean of zero. As was the case in Blitz and Modeste's (2015) analysis of the validation sample, the administrator means on each subdomain are consistently higher than teacher means. This difference shows that administrators are much more confident about the presence of a particular leadership practice than teachers (See Appendix E). The subdomain averages

for administrators are always positive (above the mean), but the subdomain averages for teachers are negative (below the mean). When comparing the means of these two groups with a t-test, I started with administrators first, as administrators are the focus of this dissertation. I then subtracted the teacher means from the administrator means. Subtracting a positive number from a negative number caused the mean differences to be negative in this analysis. A similar mathematic principle applies to the calculation of effect sizes. In subtracting the teacher mean from the administrator mean, effect sizes were negative. In comparison, all of the effect sizes were negative because teachers always had a higher score. Negative effect sizes are shown in gradients of red. Being negative or positive does not affect the significance of the effect size. A negative or positive symbol with z-scores indicates the direction from the mean.

I have created a number line in Figure 5 to illustrate the mathematical principle around positive and negative means. The z-scores range from -4 to 4. However, means for both teachers and administrators range from -1 to 1. I plotted the means for *school-wide focus on learning for teachers and administrators* (1.1) on the number line below. Administrators have a subdomain mean of .407, and teachers have a subdomain mean of -.027. To calculate the mean difference, the teacher scores of .407 was subtracted from the administrator score of -.027 to create a difference of -.434 and an effect size of -.437. The mean difference is negative because it is below the mean.

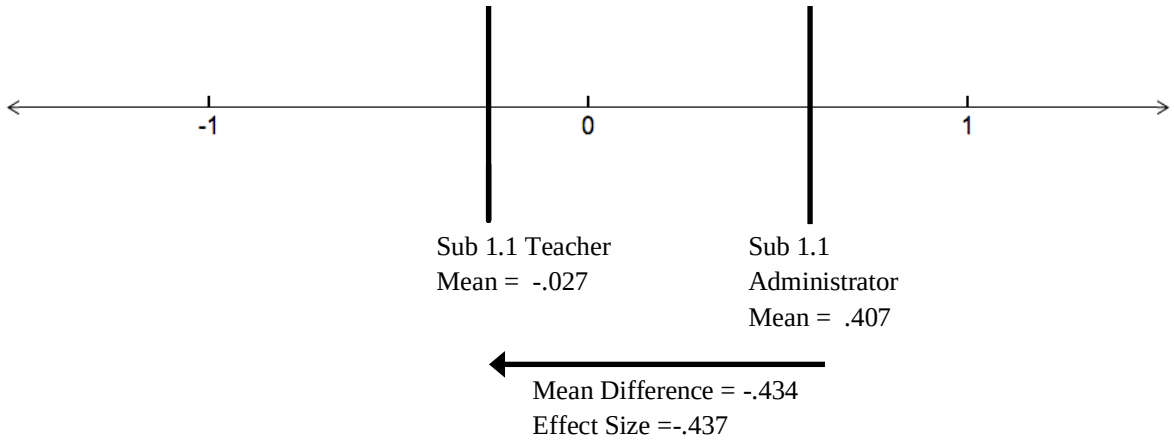


Figure 5. Number Line to Illustrate Z-scores Mean Differences.

In the first comparison, which includes the whole sample, subtracting the teacher score from the administrator score created negative mean differences. This is not the case for the other analyses in this dissertation study. Some comparisons in the other analyses showed positive mean differences, and thus, positive effect sizes. The positive effect sizes show us that the larger mean scores were in the other direction. To use Figure 5 to illustrate, if a mean difference was positive, it would show the "arrow" was pointing in the other direction with teacher means being higher than administrator means. To distinguish between negative and positive effect sizes, gradients of red will show negative effect sizes, while gradients of blue will show positive effect sizes on each table below.

Results of the First Analysis

I ran an independent samples t-test at $p = <.05$ on each of the twenty subdomains of CALL to determine the difference between the mean scores of teachers and the mean scores of principals for each domain (See Table 6). Nineteen of the subdomains are statistically significant even when using Bonferonni's FamilyWise Error Rate adjustment of $p = <.0025$. Five subdomains have no effect or a weak effect size (0.0 to 0.2), fourteen subdomains have a small effect size (0.4 to 0.6), and one subdomain has a moderate effect size (0.4 to 0.6). All of the effect sizes are negative, meaning that the teacher score is higher in each t-test. The red gradients in Table 7 indicate how much higher that particular score is. To read the effect sizes from each t-test, I replicated Table 5 from the literature review. Heat gradients were added to each subdomain in Table 7 to illustrate the difference in effect sizes.

Table 6

Results of Independent Samples T-Test Comparing Teacher and Administrator Means

	t	Sig. (2-tailed)	Mean Difference	SE Difference	Cohen's d
Z score: Subdomain 1.1: Maintaining a school-wide focus on learning	-11.267	<.001	-.435	.039	-.437
Z score: Subdomain 1.2 Formal leaders are recognized as instructional leaders	-9.779	<.001	-.392	.040	-.393
Z score: Subdomain 1.3: Collaborative design of integrated learning plan	-3.432	.001	-.148	.043	-.148
Z score: Subdomain 1.4 Providing appropriate services for students who traditionally struggle	-1.837	.067	-.084	.046	-.084
Z score: Subdomain 2.1 Formative evaluation of student learning	-4.420	<.001	-.203	.046	-.203
Z score: Subdomain 2.2: Summative evaluation of student learning	-4.376	<.001	-.185	.042	-.185
Z score: Subdomain 2.3 Formative evaluation of teaching	-8.347	<.001	-.371	.044	-.373
Z score: Subdomain 2.4 Summative evaluation of teaching	-9.119	<.001	-.367	.040	-.368
Z score: Subdomain 3.1 Collaborative school-wide focus on problems of teaching and learning	-6.325	<.001	-.275	.044	-.276
Z score: Subdomain 3.2 Professional learning	-6.305	<.001	-.272	.043	-.273
Z score: Subdomain 3.3 Socially distributed leadership	-5.605	<.001	-.240	.043	-.241
Z score: Subdomain 3.4 Coaching and Mentoring	-3.919	<.001	-.183	.047	-.183
Z score: Subdomain 4.1 Personnel practices	-7.862	<.001	-.331	.042	-.332
Z score: Subdomain 4.2 Structuring and maintaining time	-5.023	<.001	-.212	.042	-.212
Z score: Subdomain 4.3 School resources are focused on student learning	-8.211	<.001	-.353	.043	-.354
Z score: Subdomain 4.4 Integrating external expertise into the school instructional program	-4.800	<.001	-.222	.046	-.222
Z score: Subdomain 4.5 Coordinating and supervising relations with families and external communities	-7.552	<.001	-.324	.043	-.324
Z score: Subdomain 5.1 Clear, consistent, and enforced expectations for student behavior	-8.096	<.001	-.338	.042	-.339
Z score: Subdomain 5.2 Clean and safe learning environment	-3.388	.001	-.153	.045	-.154
Z score: Subdomain 5.3 Student support services provide safe haven for students who traditionally struggle	-5.476	<.001	-.242	.044	-.243

Note: Bold text indicates the one domain that is **not** statistically significant. Significant p values at $p = <.05$ with a FWER of $p = <.0025$.

Table 7

Table with Heat Gradient of Cohen's d Effect Sizes by Subdomain for Teacher vs. Administrator Means

CALL DOMAINS AND SUBDOMAINS				
1: Focus on Learning	*2: Monitoring Teaching and Learning*	*3:Building Nested Learning Communities*	*4: Acquiring and Allocating Resources*	*5: Maintaining a Safe and Effective Learning Environment*
1.1 Maintaining a school-wide focus on learning (ES= -.437)*	2.1 Formative evaluation of student learning (ES= -.203)*	3.1 Collaborative school-wide focus on problems of teaching & learning (ES= -.276)*	4.1 Personnel practices (ES= -.332)*	5.1 Clear, consistent,enforced expectations for student behavior (ES= -.339)*
1.2 Formal leaders are recognized as instructional leaders (ES= -.393)*	2.2 Summative evaluation of student learning (ES= -.185)*	3.2 Professional learning (ES= -.273)*	4.2 Structuring and maintaining time (ES= -.212)*	5.2 Safe learning environment (ES= -.154)*
1.3 Collaborative design of integrated learning plan (ES= -.148)*	2.3 Formative evaluation of teaching (ES= -.373)*	3.3 Socially distributed leadership (ES= -.241)*	4.3 School resources are focused on student learning (ES= -.354)*	5.3 Student support services for students who traditionally struggle (ES= -.243)*
1.4 Providing appropriate services for students who struggle (ES= -.084)	2.4 Summative evaluation of teaching (ES= -.368)*	3.4 Coaching and mentoring (ES= -.183)*	4.4 Integrating external expertise into school instructional program (ES= -.222)*	
			4.5 Coordinating relations with families and the external communities (ES= -.324)*	

Note: This table is presented in color. Every .10 of an increase in effect size changes the gradient of the square to be one shade darker. A larger negative effect size (deeper red) indicates that administrator means are larger. A larger positive effect size (deeper blue) indicates that teacher means are larger. Effect sizes are listed in parenthesis. *Significant p values at p = <.05 with a FWER of p = <.0025.

Tables with heat gradients are a data visualization tool that allows for differences in subdomains to be easily seen. The scale on the right side of Table 7 shows effect sizes in gradients of the color red. The effect size is calculated by subtracting the principal mean from the teacher mean. The negative effect size shows that the teacher means are lower. The white or light pink squares illustrate a weak or non-significant effect size (0.0 to 0.2). The pink and light red squares represent a small effect size (0.2 to 0.4). The dark red squares represent a moderate effect size (0.4 to 0.6). For this chart, every .10 of an increase in effect size changes the gradient of the square to be one shade darker.

Subdomains with the largest differences between teachers and administrators. The two domains in the upper left-hand corner of the table, *maintaining a school-wide focus on learning* (1.1) and *formal leaders are recognized as instructional leaders* (1.2), are the darkest red color on the map (See Table 7). The dark red indicates that they have the largest effect sizes and the greatest differences in mean scores between teachers and administrators. These subdomains show specific tasks where leaders and teachers disagree the most. The largest difference is in the area of *maintaining a school-wide focus on learning* (1.1) (MD = -.435, $ES = -.434, p = < .001$). This domain describes leaders telling and living a story that demonstrates a shared mission around student learning and touches on their ability to develop a common language around instruction at the school (Halverson & Kelly, 2017). The second-largest difference is *formal leaders are recognized as instructional leaders* (1.2), which is about a leader's ability to model active and engaged participation in professional learning, setting a clear vision for teaching and learning, and holding teachers accountable for a high level of performance (MD = -.392, $ES = -.393, p = < .001$) (Halverson & Kelly, 2017). The

two domains where there is the greatest difference between teachers and administrators are with tasks that research would associate with instructional leadership (Hallinger, 2003, 2005, 2011, 2018b; Hallinger et al., 1996; Southworth, 2002; Urick et al., 2018). These findings may affirm the difficulty of exercising instructional leadership in a way that it is seen by your followers (Bowers et al., 2017; Boyce & Bowers, 2018a; Philip Hallinger & J Murphy, 1985; Prestine & Nelson, 2005; Reich, 1988). The differences in these domains are large. Administrators rated themselves very highly on these two subdomains, indicating that they feel they are successful in telling and living a story that demonstrates a shared mission that focuses on student learning (Halverson & Kelly, 2017). However, teachers rate these two domains lower because they do not believe these leadership practices are seen in their school. This disconnect may show that instructional leadership is difficult to demonstrate. Teachers may be unaware of this aspect of principal leadership because administrators may not include them in the instructional leadership process as much as they should. Administrators may need more training in how to effectively lead instruction in a way that involves all of the teachers in the building. The large differences in these subdomains confirm other research, which also indicates that there is a difficulty for principals in exercising instructional leadership well.

Subdomains with the smallest differences between teachers and administrators. While domains with the largest difference between teachers and administrators have implications for leadership growth, domains with agreed-upon effectiveness also provide implications for how teachers and administrators prioritize leadership. In the first column at the bottom left of the table, *providing appropriate services for students who traditionally struggle* (1.4) has a white background. This lack

of color shows that this domain has a non-significant or weak effect size (MD = -.084, ES =-.084, p= .067). This domain describes how classroom teachers take ownership for the learning of all students, to what extent the school is proactive in preventing student failure, to what extent students receive differentiated instruction, and to what extent classroom and special needs teachers collaborate to develop instruction for all students (Halverson & Kelly, 2017). As the practices of *providing appropriate services for students who traditionally struggle* (1.4) are more likely to be found at the classroom level, it is not surprising as this was the smallest difference between teachers and administrators. Since teachers have more control, they generally would rate themselves higher, matching the higher scores of the administrators. The second smallest difference is for *collaborative design of an integrated learning plan* (1.3) (MD = -.148, ES =-.148, p= .001), which refers to school processes being guided by a common plan that focuses on professional development and instructional improvement (Halverson & Kelly, 2017). In Table 7, in the first column in the box second to the bottom, this subdomain is the lightest pink, indicating a weak effect size. The *collaborative design of an integrated learning plan* (1.3) deals with an essential aspect of No Child Left Behind (NCLB) and now the Every Student Succeeds Act (ESSA), both of which include creating a school-wide improvement plan. The low mean difference and non-significant effect size of *collaborative design of an integrated learning plan* (1.3) may illustrate the current political context in which administrators are now more focused on developing professional development plans to meet state and national mandates. As a collaboration between stakeholders is a mandated component of creating these plans, teachers may be more aware and involved in improvement plans than other leadership tasks. The political

context of the time may have influenced the consistent principal and teacher responses to CALL survey items. The small difference between teacher and administrator tasks on *providing appropriate services for students who traditionally struggle* (1.4) and *collaborative design of an integrated learning plan* (1.3) in these areas where teachers are more involved highlights the need for administrators to share leadership and give teachers greater ownership over school tasks. Perhaps if teachers were more familiar or involved in the leadership task, they would be able to observe the principal's leadership more clearly.

Subdomain results in the 2016-2019 CALL dataset. In reading Table 7 as a whole, frequent changes in color illustrate that effect sizes vary for each domain. The first column, *Domain 1: Focus on Learning*, has the most amount of differences in effect sizes. This range shows that there are areas where teachers greatly disagree with administrators regarding the presence of particular leadership tasks in their schools. *Domain 3: Building Professional Community*, which is the middle column of Table 7, running top to bottom has the least amount of difference in effect sizes with three of the four subdomains having a small effect size and one having a weak effect size. In this domain, the effect size or the degree to which administrators and teachers see that task displayed differently in their school is minimal. In all of the domains, the effect sizes vary, which is noteworthy as it shows the adaptive nature of leadership (Day et al., 2016; Halverson & Kelley, 2017; Leithwood & Riehl, 2005; Leithwood & Sun, 2012; V. M. Robinson et al., 2008). This analysis also highlights the need for teacher perspectives in evaluations of leadership as the administrator may think they are prioritizing a particular leadership task, but teachers may not agree (Goldring et al., 2015; Liu et al., 2014; Shoaf

et al., 2013; Tran & Bon, 2015). For example, here in Table 7, in the top right-hand corner, *clear, consistent and enforced expectations for student behavior* (5.1) has a small but significant negative effect size, indicating that teachers rate this leadership task much lower than principals (ES =-.339, p= <.001). This has significant implications as it demonstrates that administrators' enforcement of expectations may not be apparent or consistent if teachers are rating this subdomain much lower. If principal evaluation systems have the end goal of leadership growth, examining where the means of teachers and administrators do not agree, such as here in *clear, consistent and enforced expectations for student behavior* (5.1), becomes as important a learning tool as where they do agree. In examining Table 7 as a whole, the range of effect sizes provides an opportunity to discuss particular domains and the implications for those differences.

Overall, teachers and administrators have statistically significant differences in most of their ratings on subdomains. All but one of the independent samples t-tests revealed statistically significant differences at p = <.05 with a FWER of p = <.0025. Effect sizes ranged from weak to moderate, illustrating that some specific leadership practices have more disagreement between means than others. Looking at those domains with the greatest difference (*maintaining a school-wide focus on learning* (1.1) and *formal leaders are recognized as instructional leaders* (1.2)) reveals the difficultly of executing instructional leadership in a way that it is seen clearly by your followers. Conversely, the two domains with the smallest differences (*Collaborative design of integrated learning plan* (1.3) and *providing appropriate services for students who traditionally struggle* (1.4) describe practices where classroom teachers have had more direct input into the leadership behavior and generally rate those areas higher. The

presence of the largest difference *and* the smallest difference in means between teachers and administrators occurs in one domain, *Domain 1: Focus on learning*. The amount of difference in one domain illustrates the importance of examining the particular leadership for learning practices at the subdomain level, where significant differences can occur.

Results of the First Analysis Compared to Blitz and Modeste (2015) Study

The purpose of this analysis is to illustrate how teachers and administrators perceive leadership subdomains differently in two different data sets of the same survey. Examining how context affects types of leaders builds on and extends previous leadership with a different data set (Blitz & Modeste, 2015). This analysis is significant because if the premise of this study is correct, that context is a mitigating factor, the same or similar differences in means should be seen across multiple data sets when the same survey instrument is used. I compared the data from this CALL sample, which was collected between the years of 2016 and 2019, to the Blitz and Modeste (2015) validation sample of CALL data that was collected in 2012. This analysis examines the difference between the data sets to determine whether there are similarities or differences between subdomains indicating different areas for principal growth and support.

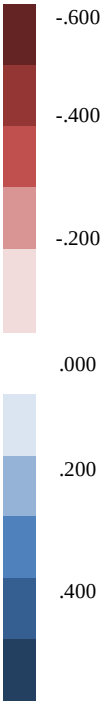
To compare the Blitz and Modeste data with the 2016-2019 CALL dataset, a table with a heat gradient is used in Tables 8 and 9 to show the effect size differences between teacher and administrator means. Tables 8 and 9 follow the same format as Table 6. There is a consistent structure in the table where darker red or blue indicates a larger difference in the responses. A red square illustrates a negative effect size or mean difference, showing that the administrators have a higher score. A blue square shows a positive effect size or mean difference, illustrating that teachers have a higher score.

Tables should be read left to right. The data from Table 8 is taken from the above analysis. Table 8 is replicated here exactly to allow for comparison between the 2016-2019 CALL dataset and the Blitz and Modeste (2015) data. The Blitz and Modeste (2015) data are not new data; it is data collected in 2012 as part of a validation sample. The only data I have from the study is what is reported in Appendix B of their study. I used mean differences as an approximate for effect sizes in Table 9 because effect sizes are not available. In 2012, there were twenty-one subdomains that included the additional subdomain of *buffering the teaching environment* (5.4) (Blitz & Modeste, 2015). The CALL team later placed questions from that subdomain into the other three subdomains in *Domain 5: Establishing a Safe and Effective Learning Environment*, and thereby eliminating *buffering the teaching environment* (5.4) (Halverson & Kelly, 2017). When two heat maps are placed next to each other, the effect sizes can be visually compared. Rank order of subdomains from the smallest difference between teachers and administrators to those that have the largest difference is shown in Appendix F. Box and whisker plots of average means for teachers and administrators by subdomain is shown in Appendix G.

Table 8

Table of Cohen's d Effect Sizes of Differences Between Teachers and Administrators by Subdomain for 2016-2019 CALL Dataset

2016-2019 CALL DATASET DOMAINS AND SUBDOMAINS				
1: Focus on Learning	*2: Monitoring Teaching & Learning*	*3:Building Nested Learning Communities*	*4: Acquiring & Allocating Resources*	*5: Maintaining Learning Environment*
1.1 School-wide focus - learning (ES = -.437)*	2.1 Formative evaluation - students (ES = -.201)*	3.1 School-wide focus - problems (ES = -.276)*	4.1 Personnel practices (ES = -.332)*	5.1 Expectations for student behavior (ES = -.339)*
1.2 Instructional leaders (ES = -.393)*	2.2 Summative evaluation - students (ES = -.185)*	3.2 Professional learning (ES = -.273)*	4.2 Structuring and maintaining time (ES = -.212)*	5.2 Safe learning environment (ES = -.154)*
1.3 Integrated learning plan (ES = -.148)*	2.3 Formative evaluation - teaching (ES = -.373)*	3.3 Socially distributed leadership (ES = -.241)*	4.3 School resources (ES = -.354)*	5.3 Student support services (ES = -.243)*
1.4 Providing appropriate services (ES = -.084)	2.4 Summative evaluation - teaching (ES = -.368)*	3.4 Coaching and mentoring (ES = -.183)*	4.4 Integrating external expertise (ES = -.222)*	
			4.5 Relations with families & external communities (ES = -.324)*	

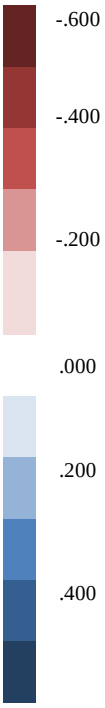


Note: *Significant p values at p = <.05 with a FWER of p = <.0025. This table is presented in color. A larger negative effect size (deeper red) indicates that administrator means are larger. A larger positive effect size (deeper blue) indicates that teacher means are larger.

Table 9

Table of Mean Differences Between Teachers and Administrators by Subdomain for Blitz and Modeste (2015) Data

BLITZ & MODESTE (2015) DOMAINS AND SUBDOMAINS				
1: Focus on Learning	*2: Monitoring Teaching & Learning*	*3:Building Nested Learning Communities*	*4: Acquiring & Allocating Resources*	*5: Maintaining Learning Environment*
1.1 School-wide focus - learning (MD = -.379)*	2.1 Formative eval-students (MD = -.056)*	3.1 School-wide focus - problems (MD = -.437)*	4.1 Personnel practices (MD = -.302)*	5.1 Expectations for student behavior (MD =-.386)*
1.2 Instructional leaders (MD = -.674)*	2.2 Summative eval-students (MD = -.053)*	3.2 Professional learning (MD = -.215)*	4.2 Structuring and maintaining time (MD=-.239)*	5.2 Safe learning environment (MD = -.502)**
1.3 Integrated learning plan (MD = -.393)*	2.3 Formative eval–teaching (MD = -.434)	3.3 Socially distributed leadership (MD = -.591)*	4.3 School resources (MD = -.419)*	5.3 Student support services (MD = -.217)*
1.4 Providing appropriate services (MD = .163)*	2.4 Summative eva -teaching (MD = -.259)	3.4 Coaching and mentoring (MD = -.225)*	4.4 Integrating external expertise (MD = -.412)*	5.4 Buffering the teaching environment (MD = -.400)*
			4.5 Relations with families & external communities (MD = -.274)*	



Note: Data and analysis using the CALL 2012 sample adapted from Blitz and Modeste (2015) *Significant p – values. This table is presented in color. A larger negative effect size (deeper red) indicates that administrator means are larger. A larger positive effect size (deeper blue) indicates that teacher means are larger.

Subdomains with the largest differences between the 2016-2019 CALL dataset and Blitz and Modeste (2015) data. As illustrated in the tables of Tables 8 and 9, a comparison of the 2016-2019 CALL dataset to the Blitz and Modeste (2015) data shows that the biggest differences in subdomains between teachers and administrators do not replicate between datasets. The differences between datasets in subdomain scores could be for several different reasons. First, it could be random error as the samples are different. Second, it could be because the 2016-2019 CALL dataset is larger, as the Blitz and Modeste (2015) data is half the size of the 2016-2019 CALL dataset. Third, the samples could be different. However, there is not enough data on the demographics of the Blitz and Modeste (2015) study to determine if the context of the two datasets is, in fact, significantly different. The significant finding in this comparison is that teachers and administrators have statistically different means in nearly all of the subdomains on both data sets, which has implications for how leaders are assessed and supported as in two samples, teachers and administrators rated leadership tasks differently. Similarly, looking at particular subdomains that have similar large differences between both groups provide an opportunity to examine the difficulty of those specific leadership tasks and why there is a difference in scores.

Blitz and Modeste (2015) have eight subdomains with a moderate effect (0.4 to 0.6), while the 2016-2019 CALL dataset only has two, as demonstrated in Tables 8 and 9. In reading Table 9, three subdomains are the darkest red, showing the greatest mean differences. *Formal leaders are recognized as instructional leaders* (1.2) (MD = -.674, p = < .001) and have the largest mean difference between teachers and administrators in the Blitz and Modeste (2015) dataset. In comparing this to Table 8, *formal leaders are*

recognized as instructional leaders (1.2) (MD = -.392, $ES = -.393, p = < .001$) has the second-largest difference in the 2016-2019 CALL dataset. This subdomain is unique in the CALL framework as it contains the only set of questions that inquire about the teacher's specific principal and the actions that the principal takes to improve teaching and learning (Blitz & Modeste, 2015). As noted by Blitz and Modeste (2015), this difference is not surprising as the principal would view these questions as a reflection on themselves alone and, thus, be more likely to inflate the rating. Another reason for the large difference in this subdomain across both the 2016-2019 CALL dataset and the Blitz and Modeste (2015) data may simply be that this leadership task is challenging to perform well.

Subdomains with the smallest differences between the 2016-2019 CALL dataset and the Blitz and Modeste (2015) data. As demonstrated in Tables 8 and 9, a comparison of the smallest differences between the 2016-2019 CALL dataset and the Blitz and Modeste (2015) data reveals the results do not replicate, except for *providing appropriate services for students who traditionally struggle* (1.4) (MD = .046, $p =$ .163). In Table 9 in the bottom left-hand corner, Subdomain 1.4 is the only subdomain that is blue. The blue indicates that the mean difference between administrators and teachers is positive, showing that the teacher mean is higher than the administrator mean. This same subdomain, *providing appropriate services for students who traditionally struggle* (1.4) (MD = -.084, ES =-.084, p= .067), is the only subdomain in the 2016-2019 CALL dataset that is not statistically significant between teachers and administrators. Of all the subdomains, *providing appropriate services for students who traditionally struggle* (1.4) focuses most on tasks that the teacher has control over as it deals with

leadership practices within the classroom. In this area, teachers would be more likely to inflate their rating as these questions are a reflection of their own leadership and not the leadership of the principal of the school. This consistent small difference may be indicating the ease of demonstrating these leadership behaviors, or it may be simply indicating that ratings were more likely to be inflated on this subdomain.

Subdomain results in comparing the 2016-2019 CALL dataset to the Blitz and Modeste (2015) data. In both the largest and the smallest differences, there is a consistent subdomain across the 2016-2019 CALL dataset and the Blitz and Modeste (2015) data. The subdomain, *formal leaders are recognized as instructional leaders* (1.2), shows a consistent large difference between both datasets. This area is unique to the survey in that it is the only question that focuses on the actions of a school's principal. Similarly, the subdomain, *providing appropriate services for students who traditionally struggle* (1.4), shows a small difference between both datasets. This subdomain is unique in that it focuses on the teacher's leadership actions within the classroom. The purpose of the formative nature of the CALL is for survey participants to self-reflect and improve their practice. Each multiple-choice survey question provides a range of answers that can guide a school community from one level of practice to the next (Halverson & Kelly, 2017). These areas seem to illustrate rating inflation where both administrators and teachers rated themselves high in areas where they had primary responsibility.

Comparison of 2016-2019 CALL Data by School Level

The second research question asks: **To what extent do elementary and secondary school administrator perspectives differ based on ratings within the CALL assessment? To what extent do elementary teacher and administrator perspectives differ based on ratings within the CALL assessment? To what extent do secondary teacher and administrator perspectives differ based on ratings within the CALL assessment?** In these analyses, the extent of statistically significant difference subdomain means is compared by the school level. I separated this question into two separate analyses. First, I ran twenty independent samples t-tests on only the administrator subdomain means to determine to what extent there is a difference in means between elementary and secondary administrators. Second, I returned to the entire dataset to compare elementary teacher and administrator means to secondary teacher and administrator means through running two sets of t-tests.

Results from Elementary and Secondary Administrators Comparison

After separating the administrator survey responses from the teacher responses, I ran twenty independent t-tests, one on each of the twenty subdomains, to compare the means of the elementary administrators against the means of the secondary leaders. The tests were run at $p = <.05$ with a Bonferonni adjustment of $p = <.0025$. The number of administrators varied in each t-test as the number of participants for each subdomain varies since, as was stated previously, all questions were not required to be answered by all participants, and many questions contained "I don't know" responses. Following the recommendation of the CALL team, I coded "I don't know" responses as missing (M.

Blitz, personal communication, December 23, 2019). The number of elementary administrators ranged from 300 to 333, and the number of secondary administrators ranged from 186 to 201. In Appendix H, I list the descriptive statistics by subdomain for both elementary and secondary administrators, including the number, mean, minimum number, maximum number, standard deviation, and standard error mean for each of the twenty subdomains.

As demonstrated in Tables 10 and 11, I find that in comparing the mean responses of elementary to secondary administrators across the 20 CALL subdomains, only one is statistically significant, *summative evaluation of student learning* (2.2), t (385.861)= 4.933; $p<0.001$. Please see Table 10 for a full list of t-test results. For box and whisker charts of elementary and secondary administrator scores, please see Appendix I. To illustrate effect sizes, a table with heat gradient is used (Table 11). This table should be read in the same way as described previously as it provides a visual representation of the data. Secondary administrators are used as the reference group for this data set. Elementary administrator means were subtracted from the secondary administrator means in the comparison. As a result, negative effect sizes will show that secondary administrators had a higher mean. Negative effect sizes will be represented in blue. Positive effect sizes will show that elementary administrators had a higher mean. Positive effect sizes will be represented in red.

Table 10

T-test Results for Elementary Administrators Subdomain Means vs. Secondary Administrators Subdomain Means

	t	df	Sig. (2-tailed)	Mean Difference	SE Difference	Cohen's d
Z score: Subdomain 1.1: Maintaining a school-wide focus on learning	-.017	440.049	.986	-.001	.075	.001
Z score: Subdomain 1.2 Formal leaders are recognized as instructional leaders	-.155	437.868	.877	-.012	.078	-.014
Z score: Subdomain 1.3: Collaborative design of integrated learning plan	-.666	484.130	.506	-.054	.082	-.057
Z score: Subdomain 1.4 Providing appropriate services for students who traditionally struggle	-2.227	404.889	.026	-.203	.091	-.203
Z score: Subdomain 2.1 Formative evaluation of student learning	2.955	410.655	.003	.267	.090	.269
Z score: Subdomain 2.2: Summative evaluation of student learning	**4.933**	**385.861**	**<.001**	**.414**	**.084**	**.457**
Z score: Subdomain 2.3 Formative evaluation of teaching	-.068	429.847	.946	-.006	.086	-.006
Z score: Subdomain 2.4 Summative evaluation of teaching	1.340	440.671	.181	.105	.078	.119
Z score: Subdomain 3.1 Collaborative school-wide focus on problems of teaching and learning	.698	442.220	.486	.059	.085	.062
Z score: Subdomain 3.2 Professional learning	-.451	406.455	.652	-.039	.086	-.041
Z score: Subdomain 3.3 Socially distributed leadership	-1.550	436.708	.122	-.130	.084	-.138
Z score: Subdomain 3.4 Coaching and Mentoring	.676	411.486	.500	.062	.092	.062
Z score: Subdomain 4.1 Personnel practices	-.571	472.555	.568	-.044	.078	-.049
Z score: Subdomain 4.2 Structuring and maintaining time	1.674	462.160	.095	.131	.078	.147
Z score: Subdomain 4.3 School resources are focused on student learning	-.638	369.835	.524	-.056	.087	-.059
Z score: Subdomain 4.4 Integrating external expertise into the school instructional program	-2.390	408.252	.017	-.217	.091	-.220
Z score: Subdomain 4.5 Coordinating and supervising relations with families and external communities	-1.223	420.374	.222	-.102	.083	-.112
Z score: Subdomain 5.1 Clear, consistent, and enforced expectations for student behavior	-1.077	451.385	.282	-.087	.081	-.095
Z score: Subdomain 5.2 Clean and safe learning environment	-1.966	420.646	.050	-.174	.089	-.178
Z score: Subdomain 5.3 Student support services provide safe haven for students who traditionally struggle	-1.903	425.282	.058	-.163	.086	-.173

Note: Bold text indicates the one subdomain that **is** statistically significant. Significant p values at p = <.05 with a FWER of p = <.0025.

Table 11

Table with Heat Gradient of Cohen's d Effect Sizes by Subdomain for Elementary Administrator Means vs. Secondary Administrator Means

CALL DOMAINS AND SUBDOMAINS				
1: Focus on Learning	*2: Monitoring Teaching and Learning*	*3:Building Nested Learning Communities*	*4: Acquiring and Allocating Resources*	*5: Maintaining a Safe and Effective Learning Environment*
1.1 Maintaining a school-wide focus on learning (ES= .001)	2.1 Formative evaluation of student learning (ES = .269)	3.1 Collaborative school-wide focus on problems of teaching and learning (ES = .062)	4.1 Personnel practices (ES = -.049)	5.1 Clear, consistent and enforced expectations for student behavior (ES =-.095)
1.2 Formal leaders are recognized as instructional leaders (ES = -.014)	2.2 Summative evaluation of student learning (ES = .457)*	3.2 Professional learning (ES = -.041)	4.2 Structuring and maintaining time (ES = .147)	5.2 Safe learning environment (ES = -.178)
1.3 Collaborative design of integrated learning plan (ES = -.057)	2.3 Formative evaluation of teaching (ES = -.006)	3.3 Socially distributed leadership (ES = -.138)	4.3 School resources are focused on student learning (ES = -.059)	5.3 Student support services for students who traditionally struggle (ES = -.173)
1.4 Providing appropriate services for students who traditionally struggle (ES= -.203)	2.4 Summative evaluation of teaching (ES = .119)	3.4 Coaching and mentoring (ES = .062)	4.4 Integrating external expertise into school instructional program (ES = -.220)	
			4.5 Coordinating and supervising relations with families and the external communities (ES = -.112)	

Note: This table is presented in color. Every .10 of an increase in effect size changes the gradient of the square to be one shade darker. A larger negative effect size (deeper red) indicates that secondary administrator means are larger. A larger positive effect size (deeper blue) indicates that elementary administrator means are larger. Effect sizes are listed in parenthesis.*Significant p values at p = <.05 with a FWER of p = <.0025.

In looking at Table 11, two findings are visually notable. First, in contrast to previous tables with heat gradients, nearly half of the boxes (nine total) are white, indicating that there is a non-significant effect size between the means of elementary and secondary administrators on these domains. However, of the boxes that do contain color, some contain gradients of blue, and some contain gradients of red. The color illustrates where there are significant differences in mean scores between elementary and secondary administrators, in six subdomains secondary administrator means are higher and in four subdomains elementary means are higher. The one statistically significant domain, *summative evaluation of student learning* (2.2) (ES = .457, p = <.001), had a moderate effect size, as illustrated in Table 11 through the dark blue square, meaning that elementary administrators have higher means. Subdomain 2.2 focuses on school-wide practices that provide feedback to students daily on their progress towards state standards (Halverson & Kelly, 2017). Between the four subdomains where there is any type of effect in this analysis, some subdomain means are higher for elementary administrators and others for secondary. In two subdomains, *providing appropriate services for students who traditionally struggle* (1.4) and *integrating external expertise into school instructional program* (4.4), the secondary administrators had higher means. Still, in *formative evaluation of student learning* (2.1) and *summative evaluation of student learning* (2.2), elementary administrators had higher means. While overall, the differences between elementary and secondary administrators are not significant, these four subdomains provide opportunities for collaboration around principal growth where administrators may be able to learn from each other.

Results from Elementary Teacher and Administrators and Secondary Teacher and Administrators Comparison

Research indicates that school level is a significant factor in how leadership is prioritized in schools (Hallinger, 2005; Leithwood & Jantzi, 2008; Nguni et al., 2006; Thoonen et al., 2011). To further test the difference in means according to the school level, I returned to the entire data sample of administrators and teachers. I split the file into two groups: elementary and secondary schools. I ran twenty independent samples t-tests to examine the difference in subdomain means between elementary teachers and administrators. At the same time, I ran twenty independent samples t-tests to examine the difference between secondary teachers and administrators. Independent samples t-tests were run at $p = <.05$ with a Bonferroni adjustment of $p = <.0025$. As seen in Table 12, the results of the t-tests indicate that ten of the subdomains are statistically significant between elementary teachers and administrators. Sixteen of the subdomains are statistically significant between secondary teachers and administrators. Thus, more domains indicate a statistically significant difference between teachers and administrators at the secondary level than at the elementary level between teachers and administrators. Quartile rankings in Appendix J list the mean differences between elementary teachers and administrators and then secondary teachers and administrators in rank order from smallest to greatest. In Appendix J, I also list the standard error difference and the upper and lower confidence intervals of difference. For box and whisker plots of the difference in means between elementary teachers and administrators and secondary teachers and administrators, please see Appendix K.

Table 12

T-test Results for Elementary Teachers vs. Administrators Means and Secondary Teachers vs. Administrators Means

	t	*df*	Sig. (2-tailed)	Mean Difference	SE Difference	Cohen's d
ELEMENTARY SCHOOLS						
Z score: Subdomain 1.1: Maintaining a school-wide focus on learning	-6.646	402.483	<.001	-.333	.050	-.341
Z score: Subdomain 1.2 Formal leaders are recognized as instructional leaders	-5.454	387.970	<.001	-.284	.052	-.291
Z score: Subdomain 1.3: Collaborative design of integrated learning plan	-3.052	384.520	.002	-.180	.059	-.169
Z score: Subdomain 1.4 Providing appropriate services for students who traditionally struggle	**-0.036**	**366.919**	**.971**	**-.002**	**.057**	**-.002**
Z score: Subdomain 2.1 Formative evaluation of student learning	**-2.888**	**359.626**	**.004**	**-.169**	**.059**	**-.173**
Z score: Subdomain 2.2: Summative evaluation of student learning	-3.779	383.834	<.001	-.193	.051	-.200
Z score: Subdomain 2.3 Formative evaluation of teaching	-5.158	359.826	<.001	-.304	.059	-.304
Z score: Subdomain 2.4 Summative evaluation of teaching	-5.935	386.009	<.001	-.313	.053	-.316
Z score: Subdomain 3.1 Collaborative school-wide focus on problems of teaching and learning	**-2.761**	**370.972**	**.006**	**-.158**	**.057**	**-.161**
Z score: Subdomain 3.2 Professional learning	-3.278	379.166	.001	-.179	.055	-.181
Z score: Subdomain 3.3 Socially distributed leadership	**-2.961**	**377.806**	**.003**	**-.166**	**.056**	**-.166**
Z score: Subdomain 3.4 Coaching and Mentoring	**-1.583**	**347.140**	**.114**	**-.096**	**.061**	**-.096**
Z score: Subdomain 4.1 Personnel practices	-4.979	351.843	<.001	-.294	.059	-.285
Z score: Subdomain 4.2 Structuring and maintaining time	**-2.399**	**351.468**	**.017**	**-.140**	**.058**	**-.136**
Z score: Subdomain 4.3 School resources are focused on student learning	-5.049	374.358	<.001	-.269	.053	-.280
Z score: Subdomain 4.4 Integrating external expertise into the school instructional program	**-1.230**	**349.526**	**.219**	**-.074**	**.060**	**-.074**
Z score: Subdomain 4.5 Coordinating and supervising relations with families and external communities	-3.298	352.924	.001	-.186	.056	-.190
Z score: Subdomain 5.1 Clear, consistent, and enforced	**-2.973**	**374.044**	**.003**	**-.165**	**.055**	**-.167**

expectations for student behavior						
Z score: Subdomain 5.2 Clean and safe learning environment	**-0.795**	**353.602**	**.427**	**-.047**	**.059**	**-.048**
Z score: Subdomain 5.3 Student support services provide safe haven for students who traditionally struggle	**-2.103**	**353.110**	**.036**	**-.123**	**.059**	**-.122**
SECONDARY SCHOOLS						
Z score: Subdomain 1.1: Maintaining a school-wide focus on learning	-9.054	232.956	<.001	-.547	.060	-.545
Z score: Subdomain 1.2 Formal leaders are recognized as instructional leaders	-8.191	229.524	<.001	-.513	.063	-.510
Z score: Subdomain 1.3: Collaborative design of integrated learning plan	**-2.043**	**224.401**	**.042**	**-.124**	**.061**	**-.135**
Z score: Subdomain 1.4 Providing appropriate services for students who traditionally struggle	**-2.902**	**216.168**	**.004**	**-.216**	**.074**	**-.206**
Z score: Subdomain 2.1 Formative evaluation of student learning	**-2.568**	**211.241**	**.011**	**-.186**	**.072**	**-.185**
Z score: Subdomain 2.2: Summative evaluation of student learning	**-1.335**	**219.519**	**.183**	**-.094**	**.070**	**-.200**
Z score: Subdomain 2.3 Formative evaluation of teaching	-6.664	211.280	<.001	-.447	.067	-.453
Z score: Subdomain 2.4 Summative evaluation of teaching	-6.558	226.290	<.001	-.405	.062	-.407
Z score: Subdomain 3.1 Collaborative school-wide focus on problems of teaching and learning	-5.919	220.903	<.001	-.393	.066	-.393
Z score: Subdomain 3.2 Professional learning	-5.427	217.705	<.001	-.382	.070	-.382
Z score: Subdomain 3.3 Socially distributed leadership	-5.270	222.239	<.001	-.349	.066	-.350
Z score: Subdomain 3.4 Coaching and Mentoring	-3.694	205.960	<.001	-.266	.072	-.271
Z score: Subdomain 4.1 Personnel practices	-6.861	221.528	<.001	-.379	.055	-.398
Z score: Subdomain 4.2 Structuring and maintaining time	-4.663	217.317	<.001	-.266	.057	-.281
Z score: Subdomain 4.3 School resources are focused on student learning	-6.332	207.690	<.001	-.461	.073	-.445
Z score: Subdomain 4.4 Integrating external expertise into the school instructional program	-5.959	207.367	<.001	-.429	.072	-.431
Z score: Subdomain 4.5 Coordinating and supervising relations with families and external communities	-7.598	213.381	<.001	-.496	.065	-.493

Z score: Subdomain 5.1 Clear, consistent, and enforced expectations for student behavior	-8.744	224.916	<.001	-.547	.063	-.556
Z score: Subdomain 5.2 Clean and safe learning environment	-4.372	213.195	<.001	-.306	.070	-.299
Z score: Subdomain 5.3 Student support services provide safe haven for students who traditionally struggle	-6.135	209.822	<.001	-.407	.066	-.414

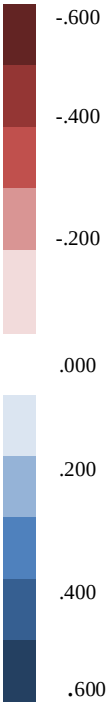
Note: Bold text indicates results that are **not** statistically significant. Significant p values at p = <.05 with a FWER of p = <.0025.

I created Tables 13 and 14 to illustrate the larger difference in effect sizes between teachers and administrators at the elementary level. For both of these tables, teacher means were subtracted from administrator means. If the effect size is negative (darker red), administrator means are higher. If the effect size is blue (darker blue), teacher means are higher. As illustrated in Tables 13 and 14, there are no blue squares because administrator means are always higher. There are three subdomains where effect sizes are similar across the elementary and secondary datasets in that teachers and administrators disagree on the extent that these leadership for learning domains are frequently practiced in their school: *school-wide focus on learning* (Subdomain 1.1; Elementary $ES = -.341, p =< .001$; Secondary $ES = -.545, p =< .001$), *formative evaluation of learning* (Subdomain 2.3; Elementary $ES = -.304, p =< .001$; Secondary $ES = -.447, p =< .001$), and *summative evaluation of learning* (Subdomain 2.4; Elementary $ES = -.313, p =< .001$; Secondary $ES = -.405, p =< .001$). These three subdomains are also the largest differences in effect sizes between teachers and administrators in the elementary comparison. The other subdomains in the comparison do not have similar effect sizes between the elementary and secondary datasets.

Table 13

Table with Heat Gradient of Cohen's d Effect Sizes of Differences Between Teachers and Administrators by Subdomain for Elementary Data

ELEMENTARY DOMAINS AND SUBDOMAINS				
1: Focus on Learning	*2: Monitoring Teaching & Learning*	*3:Building Learning Communities*	*4: Acquiring & Allocating Resources*	*5: Maintaining Learning Environment*
1.1 School-wide focus-learning (ES= -.341)*	2.1 Formative evaluation - students (ES = -.173)	3.1 School-wide focus - problems (ES = -.161)	4.1 Personnel practices (ES = -.285)*	5.1 Student behavior (ES = .167)
1.2 Instructional leaders (ES = -.291)*	2.2 Summative evaluation – students (ES = -.200)*	3.2 Professional learning (-.181)*	4.2 Structuring & maintaining time (ES = -.136)	5.2 Safe learning environment (ES = -.048)
1.3 Integrated learning plan (ES = -.169) *	2.3 Formative evaluation - teaching (ES = -.304)*	3.3 Socially distributed leadership (ES = -.166)	4.3 School resources (ES = -.280)*	5.3 Student support services (ES = -.122)
1.4 Providing appropriate services (ES= -.002)	2.4 Summative evaluation - teaching (ES =.-316)*	3.4 Coaching and mentoring (ES = -.096)	4.4 Integrating external expertise (ES = .074)	
			4.5 Relations with families & external communities (ES = -.190)*	

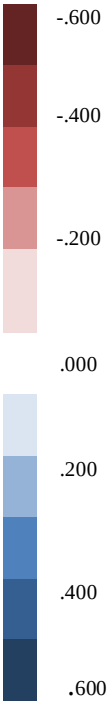


Note: <u>This table is presented in color.</u> Every .10 of an increase in effect size changes the gradient of the square to be one shade darker. A larger negative effect size (deeper red) indicates that administrator means are larger. A larger positive effect size (deeper blue) indicates that teacher means are larger.*Significant p values at p = <.05 with a FWER of p = <.0025.

Table 14

Table with Heat Gradient of Cohen's d Effect Sizes of Differences Between Teachers and Administrators by Subdomain for Secondary Data

SECONDARY DOMAINS AND SUBDOMAINS				
1: Focus on Learning	*2: Monitoring Teaching & Learning*	*3:Building Nested Learning Communities*	*4: Acquiring & Allocating Resources*	*5: Maintaining Learning Environment*
1.1 School-wide focus - learning (ES = -.545)*	2.1 Formative evaluation - students (ES = -.185)	3.1 School-wide focus – problems (ES = -.393)*	4.1 Personnel practices (ES = -.398)* **	5.1 Student behavior (ES = -.556)*
1.2 Instructional leaders (ES = -.510)*	2.2 Summative evaluation - students (ES =-.200)	3.2 Professional learning (ES = -.382)*	4.2 Structuring & maintaining time (ES = -.281)*	5.2 Safe learning environment (ES =-.299)* **
1.3 Integrated learning plan (ES = -.135)	2.3 Formative evaluation - teaching (ES = -.453)*	3.3 Socially distributed leadership (ES = -.350)*	4.3 School resources (ES = -.445)*	5.3 Student support services (ES = -414)*
1.4 Providing appropriate services (ES = -.206)	2.4 Summative evaluation - teaching (ES = -.407)*	3.4 Coaching and mentoring (ES = -.271)*	4.4 Integrating external expertise (ES = -.431)*	
			4.5 Relations with families & external communities (ES = -.493)*	



Note: This table is presented in color. Every .10 of an increase in effect size changes the gradient of the square to be one shade darker. Effect sizes are listed in parenthesis. A larger negative effect size (deeper red) indicates that administrator means are larger. A larger positive effect size (deeper blue) indicates that teacher means are larger.*Significant p values at $p = <.05$ with a FWER of $p = <.0025$. **Due to these effect sizes being .002 or less from the level above (i.e., .398, .299), the color gradient was rounded up.

Subdomains with the smallest differences in the elementary data. Overall, elementary teachers and administrators disagree less often than secondary teachers and administrators. There is alignment or agreement for elementary teachers and administrators in the 2016-2019 CALL sample on *providing appropriate support services for students who are struggling* (1.4; ES = -.002, p = .971) *coaching and mentoring* (3.4; ES = -.096, p = .114), *integrating external expertise* (4.4; ES = -.074, p = .219), and *clean and safe learning environment* (5.2; ES = -.048, p = .427). These subdomains have non-significant effect sizes and indicate agreement between elementary teachers and administrators.

Subdomains with the largest differences in the secondary data. Teachers and administrators disagree more often at the secondary level when compared to the elementary level concerning the frequency of leadership for learning practices enacted in their schools. In previous sections, the subdomains with the three highest differences were compared. In this comparison, the effect size varies so greatly that it is difficult to make a summarized comparison. In this analysis, I will compare the three subdomains that show the largest disagreement between teachers and administrators. The subdomain with the largest difference between teachers and administrators at the secondary level is *clear, consistent, and enforced expectations of student behavior* (5.1; ES = -.556, p= <.001). The questions in this section focus on teachers and administrators in the school engaging in modeling and supporting students to behave in terms of shared expectations with the school, thus providing opportunities for teachers to regularly review student discipline data and address inequities as they arise (Halverson & Kelly, 2017). This subdomain includes questions such as “how consistently do school staff members enforce

student discipline policies for all students?" The second-highest difference between secondary teachers and administrators is *maintaining a school-wide focus on learning* (1.1; ES = -.545, p= <.001). The questions in this subdomain focus on whether there is a common language for instruction across the school. Lastly, *formal leaders are recognized as instructional leaders* (1.2) has a moderate effect size in the secondary comparison (ES = -.510, p= <.001). This subdomain is where there are consistently larger differences, as shown in the previous two datasets. As previously mentioned, this subdomain encompasses leaders allocating time to model active and engaged participation in professional learning (Halverson & Kelly, 2017). The larger difference in these subdomains at the secondary level could be because, in middle and high schools, students have multiple teachers, making teachers more aware of how policies and procedures are different in classes other than their own (Hallinger, 2005, 2018). It could also be due to the sheer size of the school (Hallinger, 2005, 2018).

Overall results in comparing elementary and secondary data. The second analysis includes analyses that address an important aspect of context that was identified in the literature: school level (Hallinger, 2005; Leithwood & Jantzi, 2008; Nguni et al., 2006; Thoonen et al., 2011). When looking at the context of the school level, the significant difference occurs between the mean scores of teachers and administrators and not the mean scores between administrators at the elementary and secondary levels. Differences between the mean scores of administrators between elementary and secondary schools are not significant. Elementary and secondary administrator means were only statistically different in one domain. Only four domains even had an effect size, with three of those domains having a small effect size and one a medium effect size.

The differences between the mean scores of teachers and administrators occur more often at the secondary level than at the elementary level. When considering the frequency of leadership for learning practices enacted in their schools within the context of the school level, the difference in mean scores between teachers and administrators is more significant than between elementary and secondary administrators.

Comparison of Administrator Responses on CALL by Region

The third question in this analysis addresses the extent that there is a difference in administrator means across the twenty subdomains when controlling for region. The purpose of this analysis is to examine the question: **To what extent do administrator's ratings differ between rural, small town, suburban, and urban schools based on ratings within the CALL assessment?** To conduct this analysis, I ran a one-way ANOVA at $p = <.05$. On the CALL assessment, administrators had the option of choosing one of four categories to describe their region in relationship to region: Rural, Small Town, Suburban, and Urban. I used these four categories to run the ANOVA test. Appendix L provides the descriptive information for the responses of administrators by regional category, including mean, standard deviation, and amount of variance between groups for the ANOVA tests.

One-way ANOVA Results Comparing Administrator Means Between Regional Contexts

I conducted a one-way analysis of variance to investigate differences in means on CALL domains by region between the four categories: Rural, Small Town, Suburban, Urban (See Appendix M). ANOVA results, as presented in Appendix M, show a

significant main effect for region in six subdomains. Appendix M reports degrees of freedom between groups and within groups, the F statistic, the significance, and the effect size as measured by partial eta squared. The six domains that have significant main effects are *providing appropriate services for students who traditionally struggle* (1.4), $F(3, 396) = 5.238, p = .001$, partial $\eta^2 = .038$; *summative evaluation of student learning* (2.2), $F(3, 398) = 6.693, p < .001$, partial $\eta^2 = .048$; *summative evaluation of teaching* (2.4), $F(3, 400) = 2.655, p = .048$, partial $\eta^2 = .020$; *integrating external expertise into the school instructional program* (4.4), $F(3, 376) = 8.246, p < .001$, partial $\eta^2 = .062$; *clear, consistent, and enforced expectations of student behavior* (5.1), $F(3, 398) = 6.260, p < .001$, partial $\eta^2 = .045$; and *clean and safe learning environment* (5.2) $F(3, 384) = 2.752, p = .042$, partial $\eta^2 = .021$. Estimates of effect sizes in the other fourteen domains revealed low strength in associations. Overall, the analysis revealed that the region of the school accounts for about 5% of the variance in the means for most domain averages that is due to the region of the school. Appendix M shows these mean differences by region according to domains. Appendix N displays box and whisker charts by regional context for each domain.

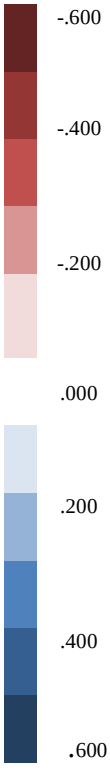
As with each of the other analyses, I use a table with a heat gradient to display the effect sizes (Table 15). The heat gradient creates a visual representation of data by shading effect sizes from least to greatest with the color red. The more saturated the color, the larger the effect. However, this graphic is different because rather than using Cohen's d, I used a partial eta squared calculation. For ANOVA's, an effect size of .10 is considered small, .25 moderate, and .40 large. In examining Table 15, you can see that it is devoid of color. This is because all of the partial eta squared analyses revealed weak

effect sizes. However, six subdomains have significant *p* values. I analyzed these *p* values using a post-hoc test to compare all group combinations and identify any significantly different pairs.

Table 15

Table with Heat Gradient of Partial Eta Squared by Subdomain for One-way ANOVA Comparing Administrator Means Between Regional Contexts

CALL DOMAINS AND SUBDOMAINS				
1: Focus on Learning	*2: Monitoring Teaching and Learning*	*3:Building Nested Learning Communities*	*4: Acquiring and Allocating Resources*	*5: Maintaining and Safe and Effective Learning Environment*
1.1 Maintaining a school-wide focus on learning (ES = .009)	2.1 Formative evaluation of student learning (ES = .020)*	3.1 Collaborative school-wide focus on problems of teaching and learning (ES = .004)	4.1 Personnel practices (ES = .014)	5.1 Clear, consistent and enforced expectations for student behavior (ES =.045)*
1.2 Formal leaders are recognized as instructional leaders (ES = .013)	2.2 Summative evaluation of student learning (ES = .048)*	3.2 Professional learning (ES = .006)	4.2 Structuring and maintaining time (ES = .007)	5.2 Safe learning environment (ES = .021)*
1.3 Collaborative design of integrated learning plan (ES = .008)	2.3 Formative evaluation of teaching (ES = .006)	3.3 Socially distributed leadership (ES = .009)	4.3 School resources are focused on student learning (ES = .001)	5.3 Student support services for students who traditionally struggle (ES =.017)
1.4 Providing services for students who traditionally struggle (ES = .038)*	2.4 Summative evaluation of teaching (ES = .020)*	3.4 Coaching and mentoring (ES = .004)	4.4 Integrating external expertise into school instructional program (ES = .062)*	
			4.5 Coordinating & supervising relations with families & external communities (ES = .001)	



Note: This table is presented in color. Every .10 of an increase in effect size changes the gradient of the square to be one shade darker. Effect sizes are listed in parenthesis.*Significant *p* values. I used Eta squared to calculate effect sizes. As it is an entirely different metric, the effect sizes are not the same scale as previous effect sizes.

Bonferonni's Post Hoc Test Results

I conducted Bonferroni's Post Hoc test to determine which groups the difference occurred for each subdomain (See Appendix O). In Appendix O, I list the mean difference, standard error, significance, and upper and lower confidence intervals for each comparison for each of the twenty subdomains. The results of this analysis, including means of each group, the mean difference, standard error, significance, and upper and lower confidence intervals, are reported below in Table 16. For brevity, only those domains with statistical significance in the means across urban categories are reported in Table 16. The results in Table 16 demonstrate that in the domains where there is a significant difference in the average score between region, this difference is always between the urban administrators and one of the other categories. It is never between one of the other categories, rural, suburban, or small town. For instance, for *integrating external expertise into the school instructional program* (4.4), the means scores between regional categories are statistically different for each comparison, Rural vs. Urban, Small Town vs. Urban, Suburban vs. Urban (See Table 16). This domain discusses the leader's ability to connect teachers to external resources to support students' learning needs. Suburban and Urban administrators have the largest difference between subdomains, as four of the six subdomains show a difference in the means between these groups.

Table 16

Comparison of Significant Subdomains with Mean Differences due to Region

	Region	Means of the Two Regions Listed**	Mean Difference	*SE*	*p*	Lower	Upper	Cohen's d
Sub 1.4 Providing appropriate services for students who traditionally struggle	Suburban vs. Urban	.533 -.043	-.577	.161	.002*	-1.005	-.149	.575
Sub 2.2: Summative evaluation of student learning	Rural vs. Urban	-.063 .306	-.369	.120	.014*	.050	.689	-.417
	Suburban vs. Urban	-.225 .306	-.531	.143	.001*	.152	.911	-.587
Sub 2.4: Summative evaluation of teaching	Suburban vs. Urban	.339 .046	.366	.135	.043*	.006	.726	-.419
Sub 4.4: Integrating external expertise into the school instructional program	Rural vs. Urban	.479 .051	.428	.134	.009*	-.784	-.073	.442
	Small town vs. Urban	.733 .051	.683	.199	.004*	-1.211	-.154	.704
	Suburban vs. Urban	.553 .051	.504	.160	.010*	-.927	-.080	.506
Sub 5.1: Clear, consistent, and enforced expectations for student behavior	Rural vs. Urban	.569 .167	.401	.123	.007*	-.728	-.075	.320
	Small town vs. Urban	.689 .167	.522	.185	.030*	-1.090	-.013	.536
Sub 5.2: Clean and safe learning environment	Small Town vs. Urban	.569 .017	.551	.203	.041*	.026	1.078	.550

* p = <.05. **Results are listed in this column as reported in the region column with the mean of the rural, small town, or suburban administrator listed first and the urban administrator mean listed second.

To keep the reports of this dissertation study's findings consistent, I created a table with a heat gradient to display the results of Bonforroni's Post Hoc test (See Table 17). While the previous heat maps looked at all twenty subdomains, this heat map only displays the six subdomains where there was a difference due to region between groups. This table shows the effect sizes between the urban administrators as significant differences only occurred between urban administrators and one of the other three groups. Suburban administrators are listed in the first column for comparison as four of the six domains showed significant differences between suburban and urban administrators. Small town administrators are in the next column and rural administrators in the last. I used Cohen's d to calculate the difference in effect sizes between these six subdomains. I started with urban administrator means and subtracted the other group's means. When the effect size is negative (deeper red), urban administrators had higher means. When the effect size is positive (deeper blue), the other group (rural, small town, suburban) had higher means. In reading Table 17, the second column shows that suburban vs. urban administrators have the largest effect sizes with three subdomain comparisons yielding large effect sizes (over .40) and one with a moderate effect size (.25-.40). In the third column, which compares small town and urban administrators, all three subdomain comparisons have a large effect. In the fourth column, which compares rural and urban administrators, two subdomains have a large effect size, and one has a moderate effect size. The rest of the effect sizes differences between the subdomain means across regions are not statistically significant. Three of effect sizes are moderate (.25-.40), and four are small (.10-.25) for those subdomains that do not have statistical significance between regional groups.

Table 17

Table with Heat Gradient of Comparison of Administrator Subdomains with Significant Mean Differences Between Urban, Suburban, Small Town, and Rural

CALL SUBDOMAINS			
1.4 Providing appropriate services for students who traditionally struggle	Suburban vs. Urban* (ES = -575)	Small Town vs. Urban (ES = .398)	Rural vs. Urban (ES = .168)
2.2: Summative evaluation of student learning	Suburban vs. Urban* (ES = -.587)	Small Town vs. Urban (ES= -.259)	Rural vs. Urban* (ES = -.417)
2.4: Summative evaluation of teaching	Suburban vs. Urban* (ES = -.419)	Small Town vs. Urban (ES= -.084)	Rural vs. Urban (ES= -.164)
4.4: Integrating external expertise into the school instructional program	Suburban vs. Urban* (ES = .506)	Small Town vs. Urban* (ES = .704)	Rural vs. Urban* (ES = .442)
5.1: Clear, consistent, and enforced expectations for student behavior	Suburban vs. Urban (ES = .355)	Small Town vs. Urban* (ES = .536)	Rural vs. Urban* (ES = .427)
5.2: Clean and safe learning environment	Suburban vs. Urban (ES =.160)	Small Town vs. Urban* (ES = .550)	Rural vs. Urban (ES =.149)

Note: This table is presented in color. Every .10 of an increase in effect size changes the gradient of the square to be one shade darker. Effect sizes are listed in parenthesis. A larger negative effect size (deeper red) indicates that urban administrator means are larger. A larger positive effect size (deeper blue) indicates that the rural, small town, or suburban means are larger. *Significant *p* values.

Though urban administrators have the most difference in means across domains, this is not because they are always lower than the mean, as illustrated in Table 17 (See Appendix L for a full list of Descriptive Statistics from this analysis). For *summative evaluation of student learning* (2.2) and *summative evaluation of teaching* (2.4), the effect size is negative, indicating that urban administrators have higher subdomain means. Conversely, in *providing appropriate services for students who traditionally struggle* (1.4), *summative evaluation of student learning* (2.2), *summative evaluation of teaching* (2.4), *integrating external expertise into the school instructional program* (4.4), *clear, consistent, and enforced expectations for student behavior* (5.1), and *clean and safe learning environment* (5.2), the effect size is positive as the rural, small town, or suburban administrators have higher subdomain means. The results show that urban administrators do not score lower on each of the six subdomains. Still, they do score distinctly different in their perceptions of their leadership practice around leadership for learning.

Overall Results in ANOVA Comparing Administrator Means Between Regional Contexts

The ANOVA test shows that 5% of the variance is accounted for by a difference in region. While the one-way ANOVA revealed that there was a difference in means, I ran a post hoc analysis to determine significant differences in six subdomains between urban administrators and the other school contexts (suburban, small town, rural). Overall, this ANOVA analysis shows differences in means in the context of region, specifically with urban administrators.

Summary of Findings

The results of this dissertation study illustrate that when measured through the CALL survey, administrators prioritized different leadership tasks in different contexts, particularly school level (elementary or secondary) and region. Overall there were five key findings. First, administrator perceptions of the frequency of leadership for learning practices are always higher than teacher scores for any given subdomain in this new 2016-2019 CALL dataset. This finding replicates and extends previous research with the CALL validation data set from 2012 (Blitz & Modeste, 2015; Bowers et al., 2017). In the present study, the largest differences indicate that principals, in comparison to teachers, more often report a focus on *maintaining a school-wide focus on learning* (1.1), *formal leaders are recognized as instructional leaders* (1.2), and *formative evaluation of teaching* (2.3). Second, the three comparisons between administrators and teachers (whole sample, elementary, secondary) had larger effect sizes across subdomains than either comparison between administrators (school level, region). Third, while this dissertation study replicates many of the findings from Blitz and Modeste's (2015) analysis of the CALL 2012 validation dataset, my analysis of the larger and more recent 2016-2019 CALL dataset indicates that in comparing the leadership for learning domains that are most different between teachers and administrators (high disagreement), only half of the CALL subdomain teacher-administrator mean differences have a similar rank order to Blitz and Modeste (2015). The difference may be due to simply different datasets (validation versus the main dataset), changing perceptions over time, or the different samples. I'll return to this issue in the discussion below. Fourth, in comparing the perceptions of the frequency of practices in leadership for learning between teachers and

administrators, elementary schools in the 2016-2019 CALL sample had many fewer disagreements across the subdomains in comparison to secondary schools, indicating perhaps a stronger alignment between teachers and administrators in leadership for learning at the elementary level. Lastly, I demonstrate that there are strong differences in the perception of leadership for learning practices between urban school administrators and administrators in other regional contexts, especially around serving the needs of struggling students, assessment, evaluation, external expertise, student discipline, and a clean environment.

Chapter V

DISCUSSION

Introduction

This chapter presents a summary of the study and the important implications drawn from the data presented in the previous chapter. First, I will give a summary of the study. Next, I will review the main findings of each analysis. Then, I discuss the implications of the major themes of the study for research, policy, and practice for principal evaluation and systems for principal growth, such as principal preparation programs. I then turn to limitations. The chapter concludes with further implications for research.

In this dissertation study, I investigated to what extent do principal leadership practices vary based on context?

The study addressed three main research questions:

1. To what extent do distributed-leadership domains differ between teacher and administrator perspectives based on ratings within the CALL assessment?
2. To what extent do elementary and secondary school leadership perspectives differ based on ratings within the CALL assessment? To what extent do elementary teacher and school leader perspectives differ based on ratings within the CALL assessment? To what extent do secondary teacher and school leader perspectives differ based on ratings within the CALL assessment?

3. To what extent do administrator's ratings on leadership for learning domains differ based on ratings within the CALL assessment between rural, small town, suburban, and urban schools?

Summary of the Study

The purpose of this dissertation study is to examine how educational leaders prioritize leadership for learning practices due to different contexts through the analysis of data from the Comprehensive Assessment of Leadership for Learning (CALL) survey. These differences will allow us to identify opportunities for preparation, support, and professional development for principals. I defined context through two different data sets, the school level, and region. I used independent samples t-tests in the first two analyses and a one-way ANOVA for the third.

The results of this dissertation study illustrate that when measured through the CALL survey, administrators prioritized different leadership tasks in different contexts, particularly school level (elementary or secondary) and region. Overall there were five key findings. First, administrator perceptions of the frequency of leadership for learning practices are always higher than teacher scores for any given subdomain in this new 2016-2019 CALL dataset. This finding replicates and extends previous research with the CALL validation data set from 2012 (Blitz & Modeste, 2015; Bowers et al., 2017). Second, comparisons of teacher and administrator scores in any of the analyses yielded larger effect sizes than comparisons between different groups of administrators. Third, while the present study replicates many of the findings from Blitz and Modeste's (2015) analysis of the CALL 2012 validation dataset, my analysis of the larger and more recent

2016-2019 CALL dataset indicates that in comparing the leadership for learning domains that are most different between teachers and administrators (large disagreement), only half of the CALL subdomain teacher-administrator mean differences have a similar rank order to Blitz and Modeste (2015). Fourth, in comparing the perceptions of the frequency of practices in leadership for learning between teachers and administrators, elementary schools in the 2016-2019 CALL sample had fewer disagreements across the subdomains in comparison to secondary schools, perhaps indicating a stronger alignment between teachers and administrators in leadership for learning at the elementary level. Lastly, I demonstrate that there are great differences in the perception of leadership for learning practices between urban school administrators and administrators in other regional contexts, especially around serving the needs of struggling students, assessment, evaluation and mentoring of teachers, and student discipline. The specifics of these implications are expanded upon below.

Application of Findings for Research, Policy, and Practice

Through this dissertation study, several specific contributions to theory and practice are gained. This section discusses the implications of the following:

a) the large difference in mean scores on most of the subdomains between teachers and administrators,

b) the persistence of some particular subdomains,

c) the stronger alignment between teachers and administrators at the elementary level, and

d) the large differences in the perception of leadership for learning practices between urban administrators and the other regional contexts.

Differences between Teachers and Administrators

There has been much research on the indirect effects of principal leadership as measured through their actions that influence school and classroom conditions and teacher-perceived student outcomes. This dissertation study contributes to the research by examining principal's perceptions of their own leadership (Agasisti et al., 2019; Hallinger, 2003, 2005; Hallinger & Heck, 1996; Leithwood, 1994; Leithwood & Jantzi, 2005; Southworth, 2002; Urick & Bowers, 2011, 2014a, 2014b, 2019). Even with the self-reported nature of the data, leadership perceptions around particular domains varied greatly between teachers and administrators.

Significant to this research is the idea of how principals' perceptions of their leadership match with teachers' perceptions of their leadership. If the purpose of leadership is to make an impact on its followers, how leaders and teachers perceive the same domains or skills is important to examine the effectiveness of leadership in those particular areas (Bowers, 2020; Bowers et al., 2017; Boyce & Bowers, 2018a; Reich, 1988; Urick & Bowers, 2019). In this dissertation study, teacher and administrator means had larger effect sizes in nineteen of the twenty subdomains in the entire sample, sixteen subdomains between secondary teachers and administrators, and ten subdomains between elementary teachers and administrators. The differences between these groups aid in our understanding of how leadership is perceived. The results show that administrators and teachers do not always agree on the frequency of the demonstration of the leadership for learning behavior (Bowers, 2020; Bowers et al., 2017). In fact, a recent study of

international survey data, the United States was unique from other countries in that administrators always rate themselves higher on leadership for learning behaviors than teachers (Bowers, 2020). While this is not the case in other countries, this consistent diffence between administrator and teacher scores replicates other research of principal leadership in the United States (Bowers, 2020; Bowers et al., 2017; Urick, 2016; Urick & Bowers, 2019).

This difference in means between teachers and administrators may suggest that administrators are not communicating their leadership in ways that can be seen and understood by their followers. For example, *formal leaders are recognized as instructional leaders* (1.2) has the largest difference in means between teachers and administrators in this 2016-2019 CALL dataset. Subdomain 1.2 is the only domain that focuses directly on the leadership of the school principal. By contrast, *providing appropriate services for students who struggle* (1.4) has the smallest difference between teachers and administrators in this 2016-2019 CALL dataset. Subdomain 1.4 is the only domain that focuses directly on the leadership of teachers in the classroom. The scores on these domains could show a tendency to inflate one's ratings, or it could show that the more a teacher is involved in leadership practice, the more likely they are to understand that behavior and recognize its presence in their performance. These large differences in means between teachers and principals may indicate the need for principals to be more proactive in communicating and involving teachers in those leadership domains. The difference in means between administrators and teachers also affirms the need for principal evaluations to involve input from not only supervisors but also teachers through systems such as a 360-degree framework (Bowers et al., 2017; Goldring et al., 2008; Liu

et al., 2015; Shoaf et al., 2013; Tran & Bon, 2015). Leaders may believe these important aspects of leadership are being enacted in a school while teachers may perceive something different. Teacher feedback will allow principals to improve in areas where teacher and principal means are divergent, resulting in growth for the leader and the school.

These large differences in subdomains between teachers and administrators echo findings in previous research, which show that principals are disconnected from teachers across many leadership for learning domains (Bowers, 2020; Bowers et al., 2017; Urick, 2016; Urick & Bowers, 2019). This difference between teacher and administrator mean scores could be due to several reasons. First, much past research has used only teacher perceptions of leadership to measure principal leadership, whereas this study and other more current research has now considered both teacher and principal perceptions of leadership (Bowers et al., 2017; Urick, 2016; Urick & Bowers, 2011, 2014b, 2019). The large difference in teacher and administrator scores in the CALL 2016-2019 dataset could be because principals understand leadership better. In their research, Urick and Bowers (2019) show that principals see instructional leadership as multi-dimensional with three main constructs from the literature while teachers only see it as one dimensional. Perhaps a similar phenomenon is happening in this dataset. Another reason for the misalignment between teacher and administrator views on many of the subdomains might be a typology alignment. The results of the CALL research done by Bowers et al. (2017) reveal that the smallest group of schools, which are schools that demonstrate high leadership for learning behaviors, have the tightest alignment between principals and teachers. As this study does not analyze any specific student achievement or school achievement factors, it

is difficult to determine whether high differences or small differences between teachers and administrators are better. For example, teachers and administrators could have high alignment but low leadership for learning practices. Additionally, perhaps administrator scores on leadership for learning practices *should* be higher than teachers on instructional leadership domains as they may be aspirational and trying to increase the leadership for learning practices in the building (Bowers et al., 2017). In considering the research of Day, Gu, Sammons (2016), perhaps administrators are demonstrating a layered approach to leadership and adapting their practices to the context. Still, teachers are slower to move to the next phase of leadership. This dissertation study provides implications for future research to explore the significance of the difference in perceptions of leadership between teachers and administrators.

In comparing the 2016-2019 dataset to the Blitz and Modeste (2015) data, my analysis indicates that of the leadership for learning domains that are the most different between teachers and administrators, only half of the CALL subdomain teacher-administrator mean differences have a similar rank order. The difference could be due to changing perceptions over time, or it could simply be that they are different data sets. Because of this, the comparison of the two groups does not have substantial implications for the understanding of how context affects leadership for learning practices. However, the difference between the groups may indicate that any two different samples of leaders would prioritize tasks differently. The difference might show the need for principal evaluations to be more aligned to district strategic goals as opposed to generic state rubrics (Close et al., 2018; Corcoran et al., 2013; Leithwood et al., 2004; Sun & Youngs, 2009). The principals' perceptions of leadership and leadership strengths varied greatly

across datasets, indicating that levers for principal growth, such as evaluation and professional development, ought to be more specific to the differentiated needs of those school leaders (Bowers et al., 2017). These differences also point to the need to move to a more developmental system of professional growth where principals build on these goals over time. The need for a developmental approach builds on prior research by Day, Gu, and Sammons (2016), who used a three year mixed methods study. They found that principals who had improved academic outcomes were able to respond to recognize the needs and motivations of others and react accordingly to the context. Day, Gu, and Sammons conclude that an educational leader's ability to sustain improvement over time is not based on a particular leadership style, but rather their ability to demonstrate specific leadership traits and behaviors at the appropriate time as determined by context. This dissertation study echoes that research by demonstrating that principals prioritize leadership for learning domains differently, perhaps responding to the needs of their context (Agasisti et al., 2019; Day et al., 2016; Hallinger, 2018a).

The Persistence of Particular Domains

While some domains were never mentioned in the narrative of this analysis, others were discussed frequently because they demonstrate repeated statistical significance in the different analyses that I conducted. These persistence differences in means on those domains may indicate the difficulty of enacting those leadership tasks. For example, *clear, consistent, and enforced expectations for student behavior* (5.1), is one of the largest differences between teachers and administrators. It is also significant between urban administrators and rural and small town administrators. Subdomain

relationships like this allow for greater research to explore how on-going training and support can be given to principals to be clearer around expectations for student behavior.

The persistent difference between teachers and administrator scores on particular subdomains may also contribute to research on theories of leadership. The two domains with the greatest differences and the largest effect sizes between teachers and administrators in the 2016-2019 CALL dataset are domains typically associated with instructional leadership. *Maintaining a school-wide focus on learning* (1.1) describes a leader's ability to develop a common language around instruction and tell a story of a school focused on learning, or in other words, be able to articulate the school's mission and direction. *Formal leaders are recognized as instructional leaders* (1.2) refers to a leader's vision for the instructional program (teaching and learning) and holding teachers accountable for high levels of performance. Instructional leadership is defined by shaping the school's mission and direction, managing the instructional program, and promoting a positive school climate (Hallinger, 2005, 2011, 2018a, 2018b; Urick, 2016; Urick & Bowers, 2014b; Urick et al., 2018). These two domains illustrate the essential aspects of instructional leadership. The effect size between teachers and administrators in the subdomain of *maintaining a school-wide focus on learning* (1.1) is -.437 in the entire dataset but grows to -.591 for the comparison between secondary teachers and administrators. The large difference between teachers and administrators means on these two subdomains seems to affirm earlier research that the effect size of instructional leadership appears to vary by school level and the socioeconomic status of the students since context greatly influences the type of instructional leadership the principal delivers (Hallinger, 2018). This finding of great differences between teachers and administrators

in the area of instructional leadership also affirms the findings of Boyce and Bowers (2018). They examine teacher data from the National Center for Education Statistics 2011–12 Schools and Staffing Survey (SASS) to find that individual teacher perceive instructional leadership as consisting of six different factors. However, collectively teachers perceive instructional leadership differently. In other words, teachers are more aware of whether or not a particular practice is occurring in the school than of the function that the leadership task is fulfilling (Boyce & Bowers, 2018b). The large differences in the 2016-2019 CALL dataset between teachers and administrators may show the administrators perceive instructional leadership as a whole construct. In contrast, teachers focus on whether or not they have seen the leadership for learning practice enacted.

Domains with consistent small differences between teachers and administrators also have implications for policy, specifically principal evaluation methods. For example, *providing appropriate services to students who traditionally struggle* (1.4) lacks statistical significance in now three CALL studies (See Figure 1 for CALL studies). In Blitz and Modeste's (2015) study, *providing appropriate services to students who traditionally struggle* (1.4) was the only subdomain in which teachers' means were higher than administrator means. Blitz and Modeste hypothesize that this is because the domain deals mainly with the teacher's own classroom, and therefore, is the domain in which teachers have the most control over the leadership practices. In Bowers et al. (2017), the typology of leaders is examined through a Latent Class Analysis. One class, the low leadership for learning teachers, has a high response rate to this item, more so than the other two teacher typologies, showing that for low leadership for learning teachers,

providing appropriate services to students who traditionally struggle (1.4) is a top priority. In this dissertation study, *providing appropriate services to students who traditionally struggle* (1.4) is the only subdomain that is not statistically significant between the teachers and administrators in the whole group sample. The fact that this domain is largely under the teacher's control may be the cause. The high teacher mean scores may demonstrate the bias of self-reported data as an individual may most often rate themselves more highly than others. The small difference between teacher and administrator tasks on *providing appropriate services for students who traditionally struggle* (1.4) in an area where teachers are more involved could also highlight the need for administrators to share leadership and give teachers greater ownership over school tasks. Perhaps if teachers were more familiar or involved in the leadership task, they would be able to observe the principal's leadership more clearly. The involvement of teachers in the leadership tasks affirms previous research that illustrates differences in principal's perceptions of leadership due to the degree to which teachers were incorporated into the leadership tasks by their principal (Urick, 2016). Once again, this particular domain illustrates the need for 360-degree frameworks that incorporate multiple perspectives on the leadership of the building through responses from teachers and principal supervisors since a leader may not fully see themselves (Bowers et al., 2017). These 360-degree frameworks may also highlight areas where teachers can be more involved in the leadership for learning practices.

Principal Leadership and Context: School Level

The second analysis addresses the context of the school level, which was identified as important in the literature (Hallinger, 2005; Leithwood & Jantzi, 2008;

Nguni et al., 2006; Thoonen et al., 2011). There was more difference in means between the groups of teachers and administrators than between elementary and secondary administrators, indicating the school level was not a significant aspect of context for administrators. However, the specific domains that have an effect size between elementary and secondary administrators do hold implications for practice. Elementary school administrators rate themselves higher in both *formative and summative evaluation of student learning* (2.1, 2.2). Secondary administrators assess themselves higher in terms *providing services for students who traditionally struggle* and *integrating external expertise* (1.4, 4.4). Since one group does not rate themselves consistently higher than the other, both groups might benefit from the opportunity to discuss best practices around those particular domains. State offices might consider bringing elementary and secondary administrators together to compare methods of leadership around these four domains with the goal of improvement (Bowers, 2010, 2015; Bowers et al., 2018; Ni et al., 2016).

In examining the entire data set of administrators and teachers, elementary schools in the 2016-2019 CALL sample had fewer disagreements across the subdomains in comparison to secondary schools, perhaps indicating a more substantial alignment between teachers and administrators in leadership for learning at the elementary level. The greater difference at the secondary level could affirm the research on instructional leadership that finds in high schools, teachers often have more content expertise in their particular content area than principals. This range of content makes managing the instructional program more difficult for principals to lead (Hallinger, 2005, 2018). Because the content is so specialized, teachers may be more likely to discredit principals' advice as they may view themselves as more of the expert.

This difference could also be due to the sheer size of the school (Hallinger, 2005, 2018). Perhaps in high schools, which are typically larger, it is more difficult for teachers to observe multiple aspects of principal leadership, making secondary perspectives less aligned. Organizational theory would tell us that the larger the organization, the more bureaucracy (Cusick, 1992; Perrow, 1973). As high schools are typically larger, there could be a greater amount of bureaucracy in the form of multiple administrators, such as assistant principals and department chairs. In rating the leadership of the building in an elementary school, teachers are more likely to be rating a single leader, whereas in a secondary school, teachers would be rating the leadership of a team. The span of control is much larger in a high school, which may account for the difference in the scores between elementary and secondary teachers and administrators (Halverson & Kelley, 2017; Raiford, 2004). Similarly, organizational theory would tell us that organizations are complex with many sources of ambiguity (Bolman & Deal, 2017). A larger school, such as a high school, would increase in complexity. Teachers are more departmentalized by subjects due to the unique content they teach. This compartmentalization might cause them to rate the leadership of the building lower because they are not seeing the leadership as a whole. Thus, while there is more statstitically significant difference between secondary teachers and administrators than elementary teachers and administrators, it is difficult to determine if this is because their views are truly different or if the results are merely an evidence of organizational theory due to the larger size of high schools and the greater specialization.

As mentioned in the methods, the CALL team separates the elementary and secondary surveys. The findings of the present study affirm this separation as ratings

between teachers and administrators are different between elementary and secondary groups. If states move to a 360-model of principal evaluation, they would be wise to follow the lead of the CALL team by differentiating questions between elementary administrators and teachers and between secondary administrators and teachers. Similarly, the finding of greater effect sizes between teachers and administrators at the secondary level affirms the previous research of Bowers et al. (2017), which showed teachers with low leadership for learning practices to often be at the high school level. As elementary and secondary administrators in this 2016-2019 CALL dataset score similarly on all of the subdomains, perhaps the greater difference in means between teachers and administrators at the secondary level is because secondary teachers exhibit lower leadership for learning practices. Though this study does not explore the reasons for the differences between school levels, the differences and similarities between school-level groups of teachers and administrators provide specific strengths and weaknesses of each school level for districts to focus on in building capacity. The differences illustrate the need for districts to provide differentiated professional development and support by grade level to administrators and their teachers.

Principal Leadership and Context: Region

The final analysis examines how the context of region affects principal leadership. The analysis identified that there are strong differences in the perception of leadership for learning practices between urban school administrators and administrators in other contexts. This dissertation study affirms research that shows that urban schools are unique to other contexts (Bowers et al., 2018; Bowers & White, 2014; Fuller et al., 2015; Greenlee & Brown Jr., 2009; Horsford et al., 2018; Hughes, Matt, & O'Reilly, 2015;

Hunter & Donahoo, 2003; R. Ingersoll & May, 2016; Johnson et al., 2012; White et al., 2012). Specifically, the six subdomains where urban administrators rate themselves differently have significant implications. Hallinger (2018) writes, "scholars have conducted empirical studies of leadership in urban and rural schools for close to 100 years. Yet, there have been no systematic reviews that carefully analyze and distinguish the types of leadership strategies that characterize these different contexts." This dissertation study answers this need by providing evidence of six leadership strategies that characterize these contexts. Two subdomains deal with the climate of the school (*clear, consistent, and enforced expectations for student behavior* (5.1) and *clean and safe learning environment* (5.2)), two subdomains deal with resourcing strategically (*providing appropriate services for students who traditionally struggle* (1.4) and *integrating external expertise into the school instructional program* (4.4), and the other two deal with summative evaluation *(summative evaluation of student learning* (2.2) and *summative evaluation of teaching* (2.4)). While there is much qualitative research on urban leadership requiring different skills, this dissertation study provides significant quantitative results that urban principals prioritize leadership differently around student discipline, resources, and summative evaluation (Khalifa et al., 2016; Theoharis, 2007).

Integrating external expertise into the school instructional program (4.4) is particularly of interest as it is statistically significant between urban administrators and each of the other three categories (Rural, Small Town, Suburban) with urban means much lower than the other three groups. This domain encompasses developing teacher expertise through articles, external professional development opportunities, district experts, and external consultants. The difference in this domain is interesting, considering the

increased grant funding that urban schools received in Title II for the express purpose of professional development. These results might indicate that though urban districts may receive increased funding for professional development, this money may not be currently used by urban principals or by those in authority above them to support teachers in the most meaningful ways (Baker, 2016).

In looking at how urban leaders prioritize leadership differently, this does not mean that their mean scores were always lower. In fact, in two domains where urban administrators have higher means deal with *summative evaluation of student learning* (2.2) and *teaching* (2.4). The difference could be due to increased testing and monitoring of urban schools after No Child Left Behind, where greater emphasis was placed on urban administrators to prove growth in summative scores for students and teachers (Hallinger, 2018; Horsford et al. 2019; Khafka, 2009). The principals from other regional contexts (rural, small town, suburban) rate themselves higher on the climate of the school (*clear, consistent, and enforced expectations for student behavior* (5.1) and *clean and safe learning environment* (5.2)) and resourcing strategically (*providing appropriate services for students who traditionally struggle* (1.4) and *integrating external expertise into the school instructional program* (4.4)). This difference could affirm research that urban schools have greater disciplinary infractions and fewer resources. Perhaps these factors cause urban administrators to rate themselves lower on these domains.

The difference in ratings for urban administrators could also be because the CALL does not fully capture the complexity of the role of urban administrator. The CALL was written to assess socially distributed leadership in a building (Halverson & Kelly, 2017). However, research suggests that collective leadership is more present in

high achieving schools (Leithwood & Mascall, 2008). In an urban environment where there are less high achieving schools, leaders could be less focused on socially distributed leadership. Similarly, issues of cultural identity and developing culturally relevant pedagogy may be a larger priority to urban leaders (Horsford et al., 2011; Khalifa et al., 2016). The CALL may not capture those cultural challenges that urban leaders face (Horsford et al., 2011).The CALL may fail to fully encapsulate the nuances of urban leadership.

The term urban can be problematic in what is being described. As illustrated by the high scores for urban leaders on summative evaluation of both teachers and administrators, this context might not be shaped by the location. Rather, it might be shaped by the environment of high-stakes accountability in meeting annual yearly progress (AYP), which is more significant in schools in urban areas (Horsford et al., 2011). Recent legislation has placed the responsibility of low-performing schools on the teachers and administrators and the metrics placed these schools mostly in urban environments (Horsford et al., 2018; Mette & Scribner, 2014). Perhaps the stress of meeting AYP causes leaders to prioritize leadership in ways that are different than those in other locations. Urban leaders are also under greater stress from increased policies, regulations, and politics (Horsford et al., 2018). This need to prioritize leadership differently in answer to the political context may be a stronger contextual factor than the location itself. The response to the political culture in urban environments is not new and may effect some subdomains more than others. As noted in recent research on culturally responsive leadership, community relations are often more controlled and rigid in an urban environment (Khalifa et al., 2016). Thus, urban as a category of region may be

encapsulating more than the location itself. The leadership differences in the six subdomains may be due to the political context of urban leadership.

These findings have significant implications for principal evaluation policy. Current principal evaluation models build on a Neo-liberal approach that views education as a free market, where teacher and principals need to be rated so that the consumers, that is parents and students, can make a decision based on the "facts" of the data (Fuller et al., 2015; Horsford et al., 2018). Most of these principal evaluation systems consist of an evaluation rubric and a data component (Close et al., 2018; Ehlert et al., 2016; Orr, 2013; Sun & Youngs, 2009). Previous literature has shown that the data used for these methods of Value-added Models is statistically unsound, making the comparison biased, particularly against low performing schools (Bowers & White, 2014; Close et al., 2018; Fuller et al., 2015; Jason A. Grissom & Susanna Loeb, 2011; Orr, 2013). The data from this dissertation study suggests that the rubric component of principal evaluation systems may also be biased because urban administrators rate themselves distinctly different from the other three groups in research-based leadership practices. These results do *not* show that urban school administrators are always underperforming, but that they are distinctly different in the way that they prioritize leadership. This research seems to indicate that the skills that needed for principals leading urban schools are different. So, if principal evaluation systems continue to remain with VAMs and state-level rubrics, this will put urban administrators at the greatest disadvantage.

The difference in urban administrator ratings on the six subdomains also provides important implications for helping administrator preparation programs better to prepare administrators for the unique challenges of urban leadership. This research indicates that

urban administrators may need a different skill set to tackle challenges that are unique to their urban context. While administrator preparation programs do not need to distinguish by the school level based on this research, a distinction between urban schools and the other regions is needed. Principal turnover is high, and urban principals are more likely to have a shorter tenure than principals in other settings (Davis, Gooden, & Bowers, 2017; Fuller et al., 2010; Levine, 2005; Young & Fuller, 2009). Perhaps administrator preparation programs that specifically focus on urban schools and the specific challenges listed in the six subdomains noted would give administrators the background in leading those schools. This targeted training may work to slow the turnover rate of principals in urban schools.

An important implication for practice is the opportunity to group urban leaders with leaders from the other three subgroups for on-going, job-embedded professional growth and development (Bowers, 2010, 2015; Bowers et al., 2018; Ni et al., 2016). Graduate schools of education are criticized for the status of underperforming schools as they trained the leaders that run them. Still, research indicates that there is a large gap of time between when most students exit their graduate educational leadership training and when they become a principal (Davis et al., 2017; Levine, 2005). Perhaps collaboration between urban administrators and administrators in other contexts would help administrators who are already in the field gain knowledge and skills from each other on how to improve on various research-based leadership tasks (Bowers, 2010, 2015; Bowers et al., 2018; Ni et al., 2016). Principals could be grouped in professional learning communities and follow protocols to create evidence-based improvement around particular domains. These groups should have the goal of implementing research-based

leadership for learning practices for school improvement (Boudett, City, & Murnane, 2013; Bowers et al., 2017; Halverson, 2010; Schildkamp, Poortman, & Handelzalts, 2016). Hargreaves and Fullan (1998, p. 106) writes that "There is no ready answer to the "how" question. Singular recipes oversimplify what it will take to bring about change in your own situation." This type of collaborative work that allows principals to process and develop strategies that would address the "how" question with critical friends could improve practice in a research-based way that would take into account each administrator's unique situation. As the CALL is a school improvement survey, this would be one very practical use for this data for county and state education offices. Networks of this type could provide real-time feedback for principal improvement that would be non-evaluative but growth-based.

Limitations

This dissertation study contains several limitations due to the methods and scope of the study. First, data were not aggregated at a school level as the aim of the study was to examine context broadly and not by pairing school responses (Blitz & Modeste, 2015). This limits the first and second analyses that compare the responses of teachers and administrators. Because the data is not nested within the school, patterns of convergence or divergence around particular domains are not able to be studied about specific leaders. I did not control for the nested dependency of the data in which teachers are nested in schools. To address this, I would need hierarchical linear models, which are outside the scope of the present study, as my purpose was to describe the differences overall across the dataset. I look to future research to examine the nested data. Similarly, this data does

not distinguish between types of leaders (Vice Principal, Department Head, Principal) or types of teachers (Content Teacher, Instructional Coach, Support Teacher), which might make a difference in responses (Blitz & Modeste, 2015). This research examines how leadership distributes across the organization with generalizing between these two groups providing a means of comparison. As the core reasoning of this dissertation study was to describe the sample overall while replicating and extending previous research, the nesting of these variables is left to future research.

Second, the difference in the amount of administrators and teachers from each state poses a threat to the validity of the findings of this study. This data is not a random sample of teachers and administrators. These responses are from school districts that chose to contract with the CALL research group to purchase the survey and the data analytics that the survey produces for school improvement. This sample was limited to participants between the years 2016 and 2019. As seen in Table 2, *Survey Participants by State as Represented in the Total Sample, Teacher Sample, and Administrator Sample*, the percentage of administrators and teachers is not equal or proportional from each state. For example, teachers from Illinois account for 14.9% of the total sample of teachers, but administrators from Illinois account for 21.3% of the total sample of administrators. This creates an imbalance in the sample that ideally would be corrected. Each state's education system has its own unique political culture that is often linked to the larger political culture of the state and may strongly influence the interactions between teachers and administrators (Louis, Thomas, & Anderson, 2010; Seashore, Leithwood, Wahlstrom, & Anderson, 2010). Some states may emphasize rationalism with polices based on comprehensive solutions to societal problems, while other states may be more

decentralized with local control and choice (Seashore et al., 2010). This difference might cause participants in the 2016-2019 CALL dataset to rate leadership practices differently. The strength of the teachers' union in a particular state may cause teachers in one state to rate their principal much lower than teachers in another state because of a political distrust of administration. This imbalance in percentages of teachers and administrators may account for the difference in the first analysis.

Because there are more teachers than administrators than teachers in a given school building, the states in Table 2 that have a larger percentage of administrators in the sample than teachers are a larger concern to validity. There are two states that have a greater percentage of administrators by over 6%, Indiana and Illinois. Their political cultures are discussed here to illustrate this validity concern. First, Indiana has an open political culture in education with a more centralized, comprehensive approach to systems and support. There are multiple mandates with a strong emphasis on capacity building. The union is not named by the state as a key actor in determining leadership or policy (Seashore et al., 2010). This emphasis on centralization with a weaker union, might cause Indiana teachers to rate their principals more highliy in response to the political culure, as administration is seen as centralized support. By contrast, the largest teacher union in the United States, the National Eductional Association (NEA) was founded in Chicago (Carini, Powell, & Steelman, 2000). Advocacy groups in Illinois tended to focus on the priorites of individual members instead of taking into account student needs (Manna & Moffit, 2014). With a more balkanized state culture, teachers might be more likely to rate items pertaining to their own leadership higher (Subdomain 1.4), while rating administrators lower due to a lack of trust. In addition to the state

culture overall, as more of the administrators in the CALL 2016-2019 data set were from an urban region, more of the administrators in Illinois could be from Chicago. Without calculating for the culture of each individual state, the percentage of difference may skew the results, particularly of the first analysis. The differences between teachers and administrators could be entirely accounted for by the CALL 2016-2019 dataset being a non-random sample. Future research should control for nested variables by nesting teachers and administrators in schools. The research might further nest these schools within states to determine the mitigating effects of state culture.

Third, the nonrandom nature of the data creates some bias and may affect the overall means. The sample size is limited to the data provided. Because of this, the data on certain domains is skewed. When looking at kurtosis and skewness for each of the domains, some subdomains are skewed though not significantly enough for the data to be invalid (See Appendix D and E). While this sample incorporates over 8,000 responses, this number is low compared to the number of educators in the country. This nonrandom and small group of data may not be generalizable to all contexts (Schreiber & Asner-Self, 2011). Similarly, the context categories mentioned- elementary and secondary and region- can not capture all of the complexity that those categories hold. For example, some schools are K-8 and may not fit well into either group. Suburban environments are more nuanced than what the overarching title may suggest (Orfield, 2011). This also may limit the generalizability of the results of this dissertation study.

Third, these responses are self-reported, which limits their potential validity (Close et al., 2018; Schreiber & Asner-Self, 2011; White et al., 2012). However, as this dissertation study aims to examine comparisons since all survey results were self-

reported, the items of comparison for all of the data have this same limitation (Halverson & Kelly, 2017).

Fourth, many other factors may affect leadership styles that are beyond context. Specifically, racial and gender identity have a significant impact on leadership styles (Eboka, 2016; Khalifa et al., 2016; Munir & Aboidullah, 2018; Shaked, Glanz, & Gross, 2018). Race, gender, ethnicity, sexuality, ability, language, social class, and faith traditions all shape an individual's leadership and culture (Horsford et al., 2011). Cultural identity mismatch may also contribute to the difference in scores between teachers and administrators. Leaders own ability to make sense of their identity and how it relates to their context may affect their perception of the presence of specific leadership practices (Evans, 2007). However, due to the limited scope of this research, identity was not taken into account in this dissertation study.

Implications for Future Research

Implications for Research

The findings of this study replicate and extend previous research in three main ways. First, this study advances principal leadership theory in illustrating that principal leadership seems to be more adaptive than one particular theory of leadership (V. M. Robinson et al., 2008). In comparing principal subdomain means between elementary and secondary administrators and then administrators by region, scores between the groups varied across the subdomains. The difference in scores between groups provides some evidence of a typology or more layered style of leadership that is more adapative depending on contextual factors (Bowers, 2017; Day et al., 2016; Urick & Bowers,

2014b). Second, this study replicates previous research that demonstrates a large difference between teacher and principal perceptions of leadership (Bowers et al., 2017; Urick, 2016; Urick & Bowers, 2019). While I postulate different reasons for this, future research might examine the impacts of that difference through examining student achievement data in addition to the difference in means. Future research might control for nested variables by analyzing the difference in principals' and teachers' perceptions within particular schools to examine patterns of low and high differences. Third, this research provides quantitative evidence that urban principals prioritize leadership differently around student discipline, resources, and summative evaluation. However, the reasons for this difference were not explored. The difference could be simply that urban administrators have less experience as previous research shows that more experienced principals have higher perceptions of their leadership for learning behaviors (Bowers et al., 2017). The difference could also be because some districts can train aspiring principals well, while other districts do not develop and train principals successfully (Davis et al., 2017). Future research might examine the context of region further to explore the reasons why this difference in context occurs to give greater support to principals due to context.

Implications for Policy

There are several implications for policy, particularly principal evaluation systems, based on the findings of the present study. First, the difference in means of subdomain scores between teachers and administrators affirms the need for principal evaluations to involve input not only from supervisors but also from teachers through systems such as a 360-degree framework (Bowers et al., 2017; Goldring et al., 2008; Liu

et al., 2015; Shoaf et al., 2013; Tran & Bon, 2015). The great difference in response between teachers and administrators at the elementary level, as opposed to the secondary level, might indicate the need for states to differentiate questions in the 360-framework by grade level. Second, there were differences between subdomain scores in the two CALL datasets used in this study, 2016-2019 CALL dataset and Blitz and Modeste (2015) dataset. These differences show a need for principal evaluations to be more aligned to specific district-level goals as opposed to the current generic state rubrics (Close et al., 2018; Corcoran et al., 2013; Leithwood et al., 2004; Sun & Youngs, 2009). These rubrics would provide systems of professional growth where principals could build on their district and local goals over time, using a layered leadership approach where principals respond differently to their contexts over time (Agasisti et al., 2019; Day et al., 2016; Hallinger, 2018a). These rubrics would also need to differentiate by regional context. While previous research indicated statistically unsound VAM's disadvantaged urban schools, this research suggests that the rubric component of most evaluation systems may do so as well (Bowers & White, 2014; Close et al., 2018; Jason A. Grissom & Susanna Loeb, 2011; Orr, 2013). State departments of education would be wise to differentiate based on the school level and the context of region when creating 360-degree frameworks for principal evaluation that are aligned to district goals.

Even though the CALL survey is formative and the current evaluation systems are summative, the results of this study do have implications for policy (Derrington, 2018; Halverson & Kelley, 2017; Shoaf et al., 2013). The CALL can serve as an example of how aspects of summative evaluations can take a more formative approach. For example, the CALL rates leadership around specific leadership tasks that are present in the school

building as opposed to an individual leader. This approach allows for the growth of the school over time as opposed to looking at the leadership of one person (Blitz & Modeste, 2015). The CALL also fits with current theories on the need for data-driven instructional leadership by providing data on the prescence of school leadership practices (Mark H. Blitz & Modeste, 2015; Elmore, 2000). This stands in contrast to summative evaluation data which does not provide evidence of what next steps a leader or a school should take (Mark H. Blitz & Modeste, 2015; Halverson, 2010). If the purpose of principal evaluation is principal growth, incorporating these elements of formative assessment into principal evaluation systems would create summative systems that are more closely tied to principal growth by providing evidence of specific research-based behaviors that principals could improve upon over time (Corcoran et al., 2013; Grissom et al., 2015; Sun & Youngs, 2009; Tran & Bon, 2015).

The recent authorization of the Every Student Succeeds Act (ESSA) granted states the ability to set their own accountability goals within the federal framework, but few states have changed their policies on the use of principal evaluation methods to make high stakes personnel decisions (Close et al., 2018; Corcoran et al., 2013; Derrington, 2018; Fuller et al., 2015; Piro et al., 2011; Sun & Youngs, 2009). However, many researchers argue that none of the summative approaches provide measures of leadership that are valid or reliable enough to be used for high stakes decisions (Fuller et al., 2015). In current ESSA plans, 31 out of 51 states and the District of Columbia include formative data for *teachers* as a part of their evaluation systems (Close et al., 2018). Perhaps the use of formative data in principal evaluation systems will not be far behind. One possible implication of this study might be to create principal evaluation systems that eliminate

summative systems entirely. Perhaps systems instead should be formative and based on growth of the particular leader over time (Day et al., 2016; Welner & Burris, 2011).

Recent research is challenging the notion that summative principal evaluation systems need to exist separately from summative ones (Close et al., 2018; Fuller et al., 2015; Welner & Burris, 2011). As Welner and Burris (2011) write,

> The formative role can and should co-exist with the summative role in a sound evaluative system...If the evaluation system has a well- functioning formative component, few educators should require dismissal (because of improvement or voluntarily exit), instruction should improve, and student achievement should increase. (p. 2)

In a model evaluation system, formative data would be used so that principals could grow. This growth could prevent or slow principal turnover and ultimately lead to the greater academic success of the students. Other research notes that formative data can be used to further professional development for principals (Close et al., 2018). Formative systems such as the CALL might replace or be incorporated into current summative evaluation systems.

Some states have already begun the process of incorporating formative elements directly into their evaluation systems. For example, New Jersey and South Carolina are already developing principal evaluation systems that contain formative elements. New Jersey recently revised their evaluation system to promote principal growth by including formative elements, such as a self-reflection rubric, authentic evidence of practice, and discussion about practices as part of the evaluation system (Riley & Meredith, 2017). South Carolina also uses authentic evidence of practice and self-reflection. Adopting formative elements to summative systems would allow principal evaluation systems to be

more varied to match the differentiated needs of those school leaders (Bowers et al., 2017; Day et al., 2016; Welner & Burris, 2011).

States that are not currently using formative assessments are looking for opportunities to develop ways of providing more support and growth for principals. These systems of support may include formative assessment. Districts are concerned about the amount of turnover in novice principals and are looking to find new ways of evaluating principals that serve as systems for growth and support (Turnbull et al., 2016). Similarly, many districts have increased the role of a principal supervisor and are looking for ways to provide customized professional development activities, but have not yet been able to find a method to do so (Close et al., 2018; Micheaux & Parvin, 2018; Turnbull et al., 2016; Welner & Burris, 2011). Perhaps the CALL can serve as such a model. The CALL provides specific feedback around leadership practices that could be used as a way to support principals. As districts are looking for ways to incorporate growth elements, current formative models, such as the CALL can serve as a guideline (Welner & Burris, 2011).

Implications for Practice

The results of this study also contain many implications for practice. First, the difference in administrator mean scores by school level and region, provide the opportunity to group leaders with leaders from the other subgroups for the purpose of on-going, job-embedded professional growth and development (Bowers, 2010, 2015; Bowers et al., 2018; Ni et al., 2016). Collaboration between administrators from different contexts would help administrators gain knowledge and skills from each other on how to improve on various research-based leadership tasks (Bowers, 2010, 2015; Bowers et al.,

2018; Ni et al., 2016). The high differences in scores between teachers and administrators on particular subdomains provide opportunities for districts to build capacity through on-going training and support can be given to principals around those particular domains. As CALL is a school improvement survey, this is one very practical use for this data for county and state offices. Second, the difference in urban administrator mean scores on six specific subdomains provides data for administrator preparation programs on how better to prepare administrators for the unique challenge of urban leadership. As the turnover rate for principals in urban schools is high, perhaps this targeted preparation will work to slow the turnover rate by giving principals the specific skills that they need to be successful in an urban environment (Davis et al., 2017; Fuller et al., 2010; Levine, 2005; Young & Fuller, 2009).

This research also provides practical implications for the CALL team. First, the difference in mean scores and effect sizes between elementary and secondary schools affirms the CALL team's practice of conducting separate surveys by the school level. Given the large differences, the CALL team is wise in continuing this practice of differentiating questions by scfol level. Second, the large differences in mean scores in this study- between teachers and administrators, elementary and secondary levels, and urban administrators and the other regional contexts- supports the current and continued uses of CALL. The CALL was created to be an instrument for school improvement (Halverson, 2010; Halverson & Kelley, 2017). The results of the present study and the many practical implications listed above attest to the importance of this survey in providing specific, actionable feedback to state policymakers, administrator preparation programs, and districts of principals and teachers. The CALL stands as an example of

how 360-degree frameworks can provide formative assessment to principals for their growth in the leadership of their schools.

Conclusion

This dissertation study provides data on how administrators prioritize leadership practices due to context, particularly in terms of how leaders rate their performance on research-based leadership domains. The present study provides evidence that suggests that current principal evaluation systems are not fully capturing the dissimilarity in school leadership due to context. Future research will hopefully examine some of the proposed areas of improvement for principal evaluation systems, such as incorporating 360-degree feedback, school-based goals, and allowing rubrics to differentiate for region. Future research should also examine the collaboration of principals around specific leadership domains for improvement. Hopefully, this combined research will move state offices to improve principal evaluation systems to include a more research-based developmental approach for principal development.

REFERENCES

Agasisti, T., Bowers, A. J., & Soncin, M. (2019). School principals' leadership types and student achievement in the Italian context: Empirical results from a three-step latent class analysis. *Educational Management Administration & Leadership, 47*(6), 860-886.

Baker, B. D. (2016). Does money matter in education? *Albert Shanker Institute*.

Bass, B. M. (1985). *Leadership and performance beyond expectations*: Collier Macmillan.

Bass, B. M., & Avolio, B. J. (1993). Transformational leadership and organizational culture. *Public administration quarterly*, 112-121.

Bass, B. M., & Avolio, B. J. (1994). Transformational leadership and organizational culture. *The International Journal of Public Administration, 17*(3-4), 541-554.

Berkovich, I. (2016). School leaders and transformational leadership theory: time to part ways? *Journal of Educational Administration, 54*(5), 609-622.

Blase, J. J. (1987). Dimensions of effective school leadership: The teacher's perspective. *American Educational Research Journal, 24*(4), 589-610.

Blitz, M. (2012). A case study comparison of school leadership practice against the comprehensive assessment of leadership for learning (CALL) pilot results. *WCERWorking Paper*(2012-5). Retrieved from University of Wisconsin–Madison, Wisconsin Center for Education Research website: http://www.wcer.wisc.edu/publications/workingPapers/papers.php

Blitz, M. H., & Modeste, M. (2015). The Differences Across Distributed Leadership Practices by School Position According to the Comprehensive Assessment of Leadership for Learning (CALL). *Leadership and Policy in Schools, 14*(3), 341-379. doi:10.1080/15700763.2015.1024328

Blitz, M. H., & Salisbury, J. (2012). The Role of Cognitive Validity Testing to Understand Leadership Practice in the Development of CALL, the Comprehensive Assessment of Leadership for Learning (WCER Working Paper No. 2012-11). Retrieved from University of Wisconsin–Madison, Wisconsin Center for Education Research website: http://www.wcer.wisc.edu/publications/workingPapers/papers.php

Bogler, R. (2001). The influence of leadership style on teacher job satisfaction. *Educational Administration Quarterly, 37*(5), 662-683.

Bolman, L., & Deal, T. (2017). *Reframing organizations. Hoboken* (Vol. 10).

Boudett, K. P., City, E. A., & Murnane, R. J. (2013). *Data wise: A step-by-step guide to using assessment results to improve teaching and learning*: ERIC.

Bowers, A. J. (2010). Analyzing the longitudinal K-12 grading histories of entire cohorts of students: Grades, data driven decision making, dropping out and hierarchical cluster analysis. *Practical Assessment, Research, and Evaluation, 15*(1), 7.

Bowers, A. J. (2015). Site selection in school district research: A measure of effectiveness using hierarchical longitudinal growth models of performance. *School Leadership & Management, 35*(1), 39-61.

Bowers, A. J. (2017). Quantitative research methods training in education leadership and administration preparation programs as disciplined inquiry for building school improvement capacity. *Journal of Research on Leadership Education, 12*(1), 72-96.

Bowers, A. J. (2020). Examining a congruency-typology model of leadership for learning using two-level latent class analysis with TALIS 2018.

Bowers, A. J., Blitz, M., Modeste, M., Salisbury, J., & Halverson, R. (2017). How leaders agree with teachers in schools on measures of leadership practice: A two-level latent class analysis of the comprehensive assessment of leadership for learning.

Bowers, A. J., Ni, X., & Esswein, J. (2018). Using hierarchical growth modeling to promote district systematic improvement in Ohio and Texas. *Leading holistically: How schools, districts, and states improve systemically*, 77-100.

Bowers, A. J., & White, B. (2014). Do principal preparation and teacher qualifications influence different types of school growth trajectories in Illinois? A growth mixture model analysis. *Journal of Educational Administration, 52*(5), 705-736.

Boyatzis, R. E., Boyatzis, R., & McKee, A. (2005). *Resonant leadership: Renewing yourself and connecting with others through mindfulness, hope, and compassion*: Harvard Business Press.

Boyce, J., & Bowers, A. J. (2018a). Different levels of leadership for learning: investigating differences between teachers individually and collectively using multilevel factor analysis of the 2011-2012 Schools and Staffing Survey. *International Journal of Leadership in Education, 21*(2), 197-225. doi:10.1080/13603124.2016.1139187

Boyce, J., & Bowers, A. J. (2018b). Toward an evolving conceptualization of instructional leadership as leadership for learning: Meta-narrative review of 109 quantitative studies across 25 years. *Journal of Educational Administration, 56*(2).

Brown, K., & Wynn, S. (2009). Finding, supporting, and keeping: The role of the principal in teacher retention issues. *Leadership and Policy in Schools, 8*(1), 37-63.

Bryk, A. S., Sebring, P. B., Allensworth, E., Easton, J. Q., & Luppescu, S. (2010). *Organizing schools for improvement: Lessons from Chicago*: University of Chicago Press.

Burns, J. M. (1978). *Leadership*. New York: Harper & Row.

Byrne-Jiménez, M., & Orr, M. T. (2012). Thinking in Three Dimensions. *Journal of Cases in Educational Leadership*, *15*(3), 33-46. doi:10.1177/1555458912447842

Camburn, E., & Salisbury, J. (2012). *Examining the Validity and Reliability of the Comprehensive Assessment of Leadership for Learning (CALL) Formative Leadership Assessment Tool: Pilot Study Results*. Paper presented at the Annual Conference of the American Education Research Association, Vancouver BC.

Carini, R., Powell, B., & Steelman, L. C. (2000). Do teacher unions hinder educational performance?: Lessons learned from State SAT and ACT scores. *Harvard Educational Review, 70*(4), 437-467.

City, E. A., Elmore, R. F., Fiarman, S. E., & Teitel, L. (2009). *Instructional rounds in education*: Cambridge, MA: Harvard Education Press.

Close, K., Amrein-Beardsley, A., & Collins, C. (2018). State-Level Assessments and Teacher Evaluation Systems after the Passage of the Every Student Succeeds Act: Some Steps in the Right Direction. *National Education Policy Center*.

Cohen, J. (1992). A power primer. *Psychological bulletin, 112*(1), 155.

Cohen, J., Cohen, P., West, S., & Aiken, L. (2003). Applied multiple correlation/regression analysis for the social sciences. In: Hillsdale, NJ: Erlbaum.

Corcoran, A., Casserly, M., Price-Baugh, R., Walston, D., Hall, R., & Simon, C. (2013). Rethinking leadership: The changing role of principal supervisors. New York, NY: Wallace Foundation and Council of the Great City Schools. In.

Cosner, S., & Jones, M. F. (2016). Leading school-wide improvement in low-performing schools facing conditions of accountability: Key actions and considerations. *Journal of Educational Administration, 54*(1), 41-57.

Cusick, P. A. (1992). *The educational system: Its nature and logic*: McGraw-Hill College.

Davis, B. W., Gooden, M. A., & Bowers, A. J. (2017). Pathways to the principalship: An event history analysis of the careers of teachers with principal certification. *American Educational Research Journal, 54*(2), 207-240.

Day, C., Gu, Q., & Sammons, P. (2016). The impact of leadership on student outcomes: How successful school leaders use transformational and instructional strategies to make a difference. *Educational Administration Quarterly, 52*(2), 221-258.

Derrington, M. L. (2018). The Principal. In S. J. Z. J. A. Ponticell (Ed.), *The Wiley Handbook of Educational Supervision* (pp. 459-482). New York: Wiley Publishing.

Donaldson, L. (2001). *The contingency theory of organizations*: Sage.

Drago-Severson, E. (2009). *Leading adult learning: Supporting adult development in our schools*: Corwin Press.

Drago-Severson, E. (2012). *Helping educators grow: Strategies and practices for leadership development*: Harvard Education Press.

Eboka, O. C. (2016). Principals Leadership Styles and Gender Influence on Teachers Morale in Public Secondary Schools. *Journal of Education and Practice, 7*(15), 25-32.

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2016). Selecting growth measures for use in school evaluation systems: Should proportionality matter? *Educational Policy, 30*(3), 465-500.

Elmore, R. F. (2000). Building a new structure for school leadership. *Albert Shanker Institute*.

Evans, A. E. (2007). School leaders and their sensemaking about race and demographic change. *Educational Administration Quarterly, 43*(2), 159-188.

Fenton, B. (2016). Leveraging principal evaluations to scale up instructional improvement. *State Education Standard, 16*(1), 19-25.

Fuller, E. J., Hollingworth, L., & Liu, J. (2015). Evaluating state principal evaluation plans across the United States. *Journal of Research on Leadership Education, 10*(3), 164-192.

Fuller, E. J., Young, M., & Shoho, A. (2010). *Assistant principal and principal attrition.* Paper presented at the American Educational Research Association annual convention, Denver, Colorado.

Geijsel, F., Sleegers, P., Leithwood, K., & Jantzi, D. (2003). Transformational leadership effects on teachers' commitment and effort toward school reform. *Journal of Educational Administration, 41*(3), 228-256.

Goff, P., Salisbury, J., & Blitz, M. (2015). Comparing CALL and VAL-ED: An Illustrative Application of a Decision Matrix for Selecting Among Leadership Feedback Instruments. WCER Working Paper No. 2015-5. *Wisconsin Center for Education Research*.

Goldring, E., Huff, J., May, H., & Camburn, E. (2008). School context and individual characteristics: what influences principal practice? *Journal of Educational Administration, 46*(3), 332-352.

Greenlee, B., & Brown Jr., J. J. (2009). Retaining Teachers in Challenging Schools. *Education, 130*(1).

Griffith, J. (2004). Relation of principal transformational leadership to school staff job satisfaction, staff turnover, and school performance. *Journal of Educational Administration, 42*(3), 333-356.

Grissom, J. A., Kalogrides, D., & Loeb, S. (2015). Using student test scores to measure principal performance. *Educational evaluation and policy analysis, 37*(1), 3-28.

Grissom, J. A., & Loeb, S. (2011). Triangulating Principal Effectiveness. *American Educational Research Journal, 48*(5), 1091-1123. doi:10.3102/0002831211402663

Grissom, J. A., & Loeb, S. (2011). Triangulating principal effectiveness: How perspectives of parents, teachers, and assistant principals identify the central importance of managerial skills. *American Educational Research Journal, 48*(5), 1091-1123.

Gu, Q., & Johansson, O. (2013). Sustaining school performance: School contexts matter. *International Journal of Leadership in Education, 16*(3), 301-326.

Gullickson, A. R., & Howard, B. B. E. (2009). *The personnel evaluation standards: How to assess systems for evaluating educators*: Corwin Press.

Hallinger, P. (2003). Leading educational change: Reflections on the practice of instructional and transformational leadership. *Cambridge Journal of education, 33*(3), 329-352.

Hallinger, P. (2005). Instructional leadership and the school principal: A passing fancy that refuses to fade away. *Leadership and Policy in Schools, 4*(3), 221-239.

Hallinger, P. (2011). Leadership for learning: Lessons from 40 years of empirical research. *Journal of Educational Administration, 49*(2), 125-142.

Hallinger, P. (2018a). Bringing context out of the shadows of leadership. *Educational Management Administration & Leadership, 46*(1), 5-24.

Hallinger, P. (2018b). Principal Instructional Leadership: From Prescription to Theory to Practice. *The Wiley Handbook of Teaching and Learning*, 505-528.

Hallinger, P., Bickman, L., & Davis, K. (1996). School context, principal leadership, and student reading achievement. *The Elementary School Journal, 96*(5), 527-549.

Hallinger, P., & Heck, R. H. (1996). Reassessing the principal's role in school effectiveness: A review of empirical research, 1980-1995. *Educational Administration Quarterly, 32*(1), 5-44.

Hallinger, P., & Heck, R. H. (1998). Exploring the principal's contribution to school effectiveness: 1980-1995. *School Effectiveness and School Improvement, 9*(2), 157-191.

Hallinger, P., Heck, R. H., & Murphy, J. (2014). Teacher evaluation and school improvement: An analysis of the evidence. *Educational Assessment, Evaluation and Accountability, 26*(1), 5-28.

Hallinger, P., & Murphy, J. (1985). Assessing the instructional leadership behavior of principals. *Elementary School Journal, 86*, 217-248.

Hallinger, P., & Murphy, J. (1985). Assessing the instructional management behavior of principals. *The Elementary School Journal, 86*(2), 217-247.

Halverson, R. (2010). School formative feedback systems. *Peabody Journal of Education, 85*(2), 130-146.

Halverson, R., & Kelley, C. (2017). *Mapping leadership: The tasks that matter for improving teaching and learning in schools*: John Wiley & Sons.

Halverson, R., Kelley, C., & Shaw, J. (2014). A CALL for improved school leadership. *Phi Delta Kappan, 95*(6), 57-60.

Hattie, J. (2009). Visible learning. Hattie ranking: Influences and effect sizes related to student achievement. In.

Heifetz, R. A., & Linsky, M. (2002). Leadership on the Line: Staying Alive through the Dangers of Leading (Harvard Business School Press). *Boston, MA*.

Hendriks, M. A., & Scheerens, J. (2013). School leadership effects revisited: a review of empirical studies guided by indirect-effect models. *School Leadership & Management, 33*(4), 373-394.

Henig, J. R. (2013). *The end of exceptionalism in American education: The changing politics of school reform*: Harvard Education Press.

Herrmann, M., & Ross, C. (2016). Measuring Principals' Effectiveness: Results from New Jersey's First Year of Statewide Principal Evaluation. REL 2016-156. *Regional Educational Laboratory Mid-Atlantic*.

Hitt, D. H., & Tucker, P. D. (2016). Systematic review of key leader practices found to influence student achievement: A unified framework. *Review of Educational Research, 86*(2), 531-569.

Horsford, S. D., Grosland, T., & Gunn, K. M. (2011). Pedagogy of the personal and professional: Toward a framework for culturally relevant leadership. *Journal of School Leadership, 21*(4), 582-606.

Horsford, S. D., Scott, J. T., & Anderson, G. L. (2018). *The politics of education policy in an era of inequality: Possibilities for democratic schooling*: Routledge.

Hughes, A. L., Matt, J. J., & O'Reilly, F. L. (2015). Principal Support Is Imperative to the Retention of Teachers in Hard-to-Staff Schools. *Journal of Education and Training Studies, 3*(1), 129-134.

Hughes, A. L., Matt, J. J., & O`Reilly, F. L. (2015). Principal Support Is Imperative to the Retention of Teachers in Hard-to-Staff Schools. *Journal of Education and Training Studies, 3*(1), 129-134.

Hunter, R. C., & Donahoo, S. (2003). The nature of urban school politics after Brown: The need for new political knowledge, leadership, and organizational skills. *Education and Urban Society, 36*(1), 3-15.

Ingersoll, R., & May, H. (2016). Minority teacher recruitment, employment and retention: 1987 to 2013. *Learning Policy Institute, Stanford, CA*.

Ingersoll, R. M. (2001). Teacher turnover and teacher shortages: An organizational analysis. *American Educational Research Journal, 38*(3), 499-534.

Ingersoll, R. M., & Smith, T. M. (2003). The wrong solution to the teacher shortage. *Educational Leadership, 60*(8), 30-33.

Jacob, R., Goddard, R., Kim, M., Miller, R., & Goddard, Y. (2015). Exploring the causal impact of the McREL Balanced Leadership Program on leadership, principal efficacy, instructional climate, educator turnover, and student achievement. *Educational evaluation and policy analysis, 37*(3), 314-332.

Johnson, S. M., Kraft, M. A., & Papay, J. P. (2012). How context matters in high-need schools: The effects of teachers' working conditions on their professional satisfaction and their students' achievement. *Teachers College Record, 114*(10), 1-39.

Kegan, R., & Lahey, L. L. (2016). *An everyone culture: Becoming a deliberately developmental organization*: Harvard Business Review Press.

Kelley, C., & Dikkers, S. (2016). Framing feedback for school improvement around distributed leadership. *Educational Administration Quarterly, 52*(3), 392-422.

Kelley, C., Halverson, R., Camburn, E., Blitz, M., Salisbury, J., Dikkers, S., . . . Kelley, C. (2012). *The design and validation of the comprehensive assessment of leadership for learning (CALL) formative school leader assessment.* Paper presented at the American Education Research Association Conference, Vancouver, British Columbia, April.

Khalifa, M. A., Gooden, M. A., & Davis, J. E. (2016). Culturally responsive school leadership: A synthesis of the literature. *Review of Educational Research, 86*(4), 1272-1311.

Kotter, J. P., & Rathgeber, H. (2006). *Our iceberg is melting: Changing and succeeding under any conditions*: Macmillan.

Leithwood, K. (1994). Leadership for school restructuring. *Educational Administration Quarterly, 30*(4), 498-518.

Leithwood, K., & Jantzi, D. (2005). A review of transformational school leadership research 1996–2005. *Leadership and Policy in Schools, 4*(3), 177-199.

Leithwood, K., & Jantzi, D. (2008). Linking leadership to student learning: The contributions of leader efficacy. *Educational Administration Quarterly, 44*(4), 496-528.

Leithwood, K., Leonard, L., & Sharratt, L. (1998). Conditions fostering organizational learning in schools. *Educational Administration Quarterly, 34*(2), 243-276.

Leithwood, K., Louis, K. S., Anderson, S., & Wahlstrom, K. (2004). How Leadership Influences Student Learning. Review of Research. *Wallace Foundation, The*.

Leithwood, K., & Mascall, B. (2008). Collective Leadership Effects on Student Achievement. *Educational Administration Quarterly, 44*(4), 529-561. doi:10.1177/0013161x08321221

Leithwood, K., & Riehl, C. (2005). What do we already know about educational leadership. *A new agenda for research in educational leadership*, 12-27.

Leithwood, K., & Sun, J. (2012). The nature and effects of transformational school leadership: A meta-analytic review of unpublished research. *Educational Administration Quarterly, 48*(3), 387-423.

Levine, A. (2005). Educating school leaders. The education schools project. *New York: Columbia University Teachers College: Education Schools Project*.

Liu, K., Springer, J., Stuit, D., Lindsay, J., & Wan, Y. (2015). The utility of teacher and student surveys in principal evaluations: An empirical investigation. In: REL.

Louis, K. S., Thomas, E., & Anderson, S. (2010). How do states influence leadership in small districts? *Leadership and Policy in Schools, 9*(3), 328-366.

Manna, P., & Moffit, S. (2014). New education advocacy organizations in the US states. In: The Wallace Foundation.

Marks, H. M., & Louis, K. S. (1999). Teacher empowerment and the capacity for organizational learning. *Educational Administration Quarterly, 35*(5), 707-750.

Marks, H. M., & Printy, S. M. (2003). Principal leadership and school performance: An integration of transformational and instructional leadership. *Educational Administration Quarterly, 39*(3), 370-397.

Marzano, R. J., Waters, T., & McNulty, B. A. (2001). *School leadership that works: From research to results*: ASCD.

Mertler, C. A., & Reinhart, R. V. (2016). *Advanced and multivariate statistical methods: Practical application and interpretation*: Routledge.

Mette, I. M., & Scribner, J. P. (2014). Turnaround, transformational, or transactional leadership: An ethical dilemma in school reform. *Journal of Cases in Educational Leadership, 17*(4), 3-18.

Micheaux, D. J., & Parvin, J. L. (2018). PRINCIPAL EVALUATION AS A TOOL FOR GROWTH. *The Learning Professional, 39*(2), 52-55.

Min, S., Modeste, M. E., Salisbury, J., & Goff, P. T. (2016). Heeding the CALL (Comprehensive Assessment of Leadership for Learning) An inquiry into instructional collaboration among school professionals. *Journal of Educational Administration, 54*(2), 135-151.

Minarik, M. M., Thornton, B., & Perreault, G. . (2003). Systems thinking can improve teacher retention. . *The Clearing House, 76*(5), 230-234.

Modeste, M. E., Hornskov, S. B., Bjerg, H., & Kelley, C. J. (2018). School leadership practice across international policy contexts: Organizational roles and distributed leadership in Denmark and the United States. *Educational Management Administration & Leadership,* 1741143218807489.

Moolenaar, N. M., Daly, A. J., & Sleegers, P. J. (2010). Occupying the principal position: Examining relationships between transformational leadership, social network position, and schools' innovative climate. *Educational Administration Quarterly, 46*(5), 623-670.

Muijs, D., & Harris, A. (2003). Teacher leadership: A review of the literature. *Educational Management & Administration, 31*(4), 437-449.

Munir, F., & Aboidullah, M. (2018). Gender Differences in Transformational Leadership Behaviors of School Principals and Teachers' Academic Effectiveness. *Bulletin of Education and Research, 40*(1), 99-113.

Murphy, J., Elliott, S. N., Goldring, E., & Porter, A. C. (2006). Learning-Centered Leadership: A Conceptual Foundation. *Learning Sciences Institute, Vanderbilt University (NJ1)*.

Nguni, S., Sleegers, P., & Denessen, E. (2006). Transformational and transactional leadership effects on teachers' job satisfaction, organizational commitment, and organizational citizenship behavior in primary schools: The Tanzanian case. *School Effectiveness and School Improvement, 17*(2), 145-177.

Ni, X., Bowers, A. J., & Esswein, J. (2016). What counts in calculating school and district level performance index scores: A summary and analysis of academic performance index metrics across the 50 states.

Orfield, M. (2011). *American metropolitics: The new suburban reality*: Brookings Institution Press.

Orr, M. T. (2013). Review of two articles on principal effects. *National Education Policy Center*.

Perrow, C. (1973). The short and glorious history of organizational theory. *Organizational Dynamics, 2*(1), 3-15.

Piro, J., Wiemers, R., & Shutt, T. (2011). Using Student Achievement Data in Teacher and Principal Evaluations: A Policy Study. *International Journal of Educational Leadership Preparation, 6*(4), n4.

Porter, A., Goldring, E., Elliott, S., Murphy, J., Polikoff, M., & Cravens, X. (2008). Setting Performance Standards for the VAL-ED: Assessment of Principal Leadership. *Online Submission*.

Prestine, N. A., & Nelson, B. S. (2005). How can educational leaders support and promote teaching and learning? New conceptions of learning and leading in schools. *A new agenda for research in educational leadership*, 46-60.

Printy, S. M. (2008). Leadership for teacher learning: A community of practice perspective. *Educational Administration Quarterly, 44*(2), 187-226.

Printy, S. M., Marks, H. M., & Bowers, A. J. (2009). Integrated leadership: How principals and teachers share transformational and instructional influence. *Journal of School Leadership, 19*(5), 504-532.

Raiford, S. A. (2004). Relationship Between Span of Control and School Performance in Selected High and Low Achieving Public Elementary Schools in Florida.

Reich, R. (1988). The Power of Public Ideas (Cambridge, MA: Ballinger). *ReichThe Power of Public Ideas1988*.

Riley, D. L., & Meredith, J. (2017). State Efforts to Strengthen School Leadership: Insights from CCSSO Action Groups. *Policy Studies Associates, Inc.*

Robinson, V. (2011). *Student-centered leadership* (Vol. 15): John Wiley & Sons.

Robinson, V. M., Lloyd, C. A., & Rowe, K. J. (2008). The impact of leadership on student outcomes: An analysis of the differential effects of leadership types. *Educational Administration Quarterly, 44*(5), 635-674.

Schildkamp, K., Poortman, C. L., & Handelzalts, A. (2016). Data teams for school improvement. *School Effectiveness and School Improvement, 27*(2), 228-254.

Schreiber, J., & Asner-Self, K. (2011). Educational Research . Hoboken. In: John Wiley & Sons, Inc.

Seashore, K., Leithwood, K., Wahlstrom, K., & Anderson, S. (2010). Investigating the links to improved student learning: Final report of research findings.

Sebastian, J., Allensworth, E., & Stevens, D. (2014). The Influence of School Leadership on Classroom Participation: Examining Configurations of Organizational Supports. *Grantee Submission, 116*.

Sebastian, J., Allensworth, E., Wiedermann, W., Hochbein, C., & Cunningham, M. (2018). Principal leadership and school performance: An examination of instructional leadership and organizational management. *Leadership and Policy in Schools*, 1-23.

Sebastian, J., Huang, H., & Allensworth, E. (2017). Examining integrated leadership systems in high schools: connecting principal and teacher leadership to organizational processes and student outcomes. *School Effectiveness and School Improvement, 28*(3), 463-488.

Shaked, H., Glanz, J., & Gross, Z. (2018). Gender differences in instructional leadership: how male and female principals perform their instructional leadership role. *School Leadership & Management, 38*(4), 417-434.

Shoaf, L., Zigler, T., & Beebe, R. (2013). Building cohesive leadership development. *Policy Futures in Education, 11*(3), 256-266.

Sinnema, C. E., & Robinson, V. M. (2012). Goal setting in principal evaluation: Goal quality and predictors of achievement. *Leadership and Policy in Schools, 11*(2), 135-167.

Southworth, G. (2002). Instructional leadership in schools: Reflections and empirical evidence. *School Leadership & Management, 22*(1), 73-91.

Spillane, J. P. (2006). *Distributed Leadership*. San Francisco, CA: Josey Bass.

Spillane, J. P., Halverson, R., & Diamond, J. B. (2001). Investigating school leadership practice: A distributed perspective. *Educational researcher, 30*(3), 23-28.

Spillane, J. P., Halverson, R., & Diamond, J. B. (2004). Towards a theory of leadership practice: A distributed perspective. *Journal of curriculum studies, 36*(1), 3-34.

Stein, M. K., & Spillane, J. (2005). What can researchers on educational leadership learn from research on teaching? Building a bridge. *A new agenda for research in educational leadership*, 28-45.

Strayhorn, T. L. (2010). Accessing and analyzing national databases. In *Handbook of data-based decision making in education* (pp. 121-138): Routledge.

Sun, M., & Youngs, P. (2009). How does district principal evaluation affect learning-centered principal leadership? Evidence from Michigan school districts. *Leadership and Policy in Schools, 8*(4), 411-445.

Theoharis, G. (2007). Social justice educational leaders and resistance: Toward a theory of social justice leadership. *Educational Administration Quarterly, 43*(2), 221-258.

Thoonen, E. E., Sleegers, P. J., Oort, F. J., Peetsma, T. T., & Geijsel, F. P. (2011). How to improve teaching practices: The role of teacher motivation, organizational factors, and leadership practices. *Educational Administration Quarterly, 47*(3), 496-536.

Tran, H., & Bon, S. C. (2015). Assessing Multiple Stakeholders' Perceptions of an Effective Principal Evaluation System. *Education Leadership Review, 16*(2), 1-18.

Turnbull, B. J., Anderson, L. M., Riley, D. L., MacFarlane, J. R., & Aladjem, D. K. (2016). The Principal Pipeline Initiative in Action. Building a Stronger Principalship: Volume 5. *Policy Studies Associates, Inc.*

Urick, A. (2016). Examining US principal perception of multiple leadership styles used to practice shared instructional leadership. *Journal of Educational Administration, 54*(2), 152-172.

Urick, A., & Bowers, A. J. (2011). What influences principals' perceptions of academic climate? A nationally representative study of the direct effects of perception on climate. *Leadership and Policy in Schools, 10*(3), 322-348.

Urick, A., & Bowers, A. J. (2014a). The impact of principal perception on student academic climate and achievement in high school: how does it measure up? *Journal of School Leadership, 24*(2), 386-414.

Urick, A., & Bowers, A. J. (2014b). What are the different types of principals across the United States? A latent class analysis of principal perception of leadership. *Educational Administration Quarterly, 50*(1), 96-134.

Urick, A., & Bowers, A. J. (2019). Assessing international teacher and principal perceptions of instructional leadership: A multilevel factor analysis of TALIS 2008. *Leadership and Policy in Schools, 18*(3), 249-269.

Urick, A., Wilson, A. S., Ford, T. G., Frick, W. C., & Wronowski, M. L. (2018). Testing a framework of math progress indicators for ESSA: How opportunity to learn and instructional leadership matter. *Educational Administration Quarterly, 54*(3), 396-438.

Ware, H., & Kitsantas, A. (2007). Teacher and collective efficacy beliefs as predictors of professional commitment. *The Journal of Educational Research, 100*(5), 303-310.

Waters, T., Marzano, R. J., & McNulty, B. (2003). Balanced Leadership: What 30 Years of Research Tells Us about the Effect of Leadership on Student Achievement. A Working Paper.

Welner, K. G., & Burris, C. C. (2011). Letter to Secretary of Education Arne Duncan Concerning Evaluation of Teachers and Principals.

White, M. E., Makkonen, R., Vince, S., & Bailey, J. (2012). How California's Local Education Agencies Evaluate Teachers and Principals. REL Technical Brief. REL 2012-No. 023. *Regional Educational Laboratory West*.

Williams, S. M. (2015). The future of principal preparation and principal evaluation: Reflections of the current policy context for school leaders. *Journal of Research on Leadership Education, 10*(3), 222-225.

Young, M., & Fuller, E. J. (2009). The revolving door of the Texas high school principalship. *HR Exchange, Texas Association of School Boards*, 1-2.

Zembylas, M., & Papanastasiou, E. C. (2005). Modeling teacher empowerment: The role of job satisfaction. *Educational Research and Evaluation, 11*(5), 433-459.

Appendix A

CALL Survey Questions Listed by Type

Variable	Elementary (ES), Middle, High (MSHS), Both (B)	Principal (P), Teacher (T), Both (B)	Filter Question	Contains "I Don't Know" Responses
040410_b	B	B		
040410_c	B	B	X	
q010110	B	B		
q010140	B	B		
q010141	B	B		
q010142	B	B		
q010143	B	B		
q010150^	B	B		
q010160A^	B	A		
q010160T^	B	T		
q010210	B	B		
q010221	B	B		
q010222	B	B		
q010223	B	B		
q010330	B	B		
q010340	B	B		
q010350^	B	B		
q010361^	B	B		X
q010362^	B	B		X
q010363^	B	B		X
q010364^	B	B		X
q010365^	B	B		X
q010420(!)	B	B		X
q010421(!)	B	B		X
q010422ES(!)	ES	B		
q010430A	B	A		
q010430T	B	T		
q010450^	B	B		X
q010461^	B	B		
q010462^	B	B		
q010463^	B	B		
q010464^	B	B		
q010465^	B	B		
q020110(!)	B	B		X
q020120_a	B	B	X	

q020120_b	B	B			
q020121(!)	B	B			X
q020130	B	B			
q020131	B	B			
q020132HS	MSHS	B			
q020132ES	ES	B			
q020133HS	MSHS	B			
q020133ES	ES	B			
q020134	B	B			
q020135	B	B			
q020210HS	MSHS	B			
q020210ES	ES	B			
q020220(!)	B	B			X
q020230	B	B			
q020240	B	B			
q020241	B	B			
q020250	B	B			
q020310A_a	B	A	X		
q020310A_b	B	A			
q020310T_a	B	T	X		
q020310T_b	B	T			
q020320	B	B			
q020334A_a	B	A	X		
q020334A_b	B	A			
q020334T_a	B	T	X		
q020334T_b	B	T			
q020335A_a	B	A	X		
q020335A_b	B	A			
q020335T_a	B	T	X		
q020335T_b	B	T			
q020337A_a	B	A	X		
q020337A_b	B	A			
q020337T_a	B	T	X		
q020337T_b	B	T			
q020338A_a	B	A	X		
q020338A_b	B	T			
q020338T_a	B	T	X		
q020338T_b	B	T			
q020410(!)	B	B			X
q020431A	B	A			
q020431T	B	T			
q020432A	B	A			
q020432T	B	T			

q020433A	B	A			
q020433T	B	T			
q020450A	B	A			
q020450T	B	T			
q020451	B	B			
q020452	B	B			
q020460	B	B			
q020461	B	B			
q030111_b	B	B			
q030113_b	B	B			
q030114_b	B	B			
q030130A_a	B	A	X		
q030130A_b(!)	B	A			
q030130T_a	B	T	X		
q030130T_b	B	T			
q030131	B	B			
q030140^	B	B			
q030150^	B	B			X
q030220A	B	A			
q030220T	B	T			
q030230(!)	B	B			X
q030240	B	B			
q030250	B	B			
q030260^	B	B			
q030270^	B	B			
q030280^	B	B			
q030290^	B	B			
q030310	B	B	X		
q030311(!)	B	B			X
q030312(!)	B	B			X
q030313(!)	B	B			X
q030325	ES	B			
q030325OU	MSHS	B			
q030330	B	B			
q030340	B	B			
q030420	B	B	X		
q030422	B	B			X
q030424	B	B			
q030430	B	B	X		
q030432(!)	B	B			X
q030440^	B	B			
q040113(!)	B	B			X

q040115(!)	B	B		X
q040117(!)	B	B		X
q040118ES(!)	ES	B		X
q040150(!)	B	B		X
q040160	B	B		
q040231(!)	B	B		X
q040233(!)	B	B		X
q040234(!)	B	B		X
q040235ES(!)	ES	B		X
q040235HS(!)	MSHS	B		X
q040236HS(!)	MSHS	B		X
q040240^	B	B		
q040313_b(!)^	B	B		X
q040320A	B	A		
q040320T	B	T		
q040330^	ES	B		
q040340	B	B		
q040350	B	B		
q040360_b^	B	B		X
q040360_c^	B	B		X
q040410_a	B	B	X	
q040410_d	B	B		
q040420	B	B		
q040510_c	B	B		
q040510_d	B	B		
q040510_e	B	B		
q040511	B	B		
q040512	B	B		
q040520	B	B		
q040530_a	B	B	X	
q040530_b(!)	B	B		X
q040560	B	B		
q040570^	B	B		
q040580^	B	B		
q050110	B	B		
q050111_a	B	B		
q050111_b	B	B		
q050111_c	B	B		
q050111_d	B	B		
q050120	B	B		
q050121	B	B		
q050122	B	B		
q050125	B	B		

q050130	B	B			
q050140	B	B			
q050150	B	B			
q050160	B	B			
q050161	B	B			
q050162	B	B			
q050163	B	B			
q050164	B	B			
q050170	B	B			
q050250	B	B			
q050261	B	B			
q050262	B	B			
q050263	B	B			
q050267ES	ES	B			
q050268ES	ES	B			
q050271	B	B			
q050272	B	B			
q050273	B	B			
q050277ES	ES	B			
q050278ES	ES	B			
q050330_a	B	B	X		
q050330_b	B	B			
q050340	B	B			X
q050351a	B	B	X		
q050351b	B	B			
q050352a	B	B	X		
q050352b	B	B			
q050353a	B	B	X		
q050353b	B	B			
q050354_a	MSHS	B	X		
q050354_bHS	MSHS	B			
q050355a	B	B	X		
q050355b	MSHS	B			
q050360(!)	MSHS	B			X
q050370ES	ES	B			
q050370HS	MSHS	B			
Q050371ES	ES	B			
q050380HS	MSHS	B			

Appendix B

CALL Variables that Contributed to each Subdomain Mean

CALL Subdomain	Teacher Questions	Administrator Questions
Subdomain 1.1: Maintaining a school-wide focus on learning	q010110, q010140, q010141, q010142, q010143, q010150^, q010160T^	q010110, q010140, q010141, q010142, q010143, q010150^, q010160A^
Subdomain 1.2 Formal leaders are recognized as instructional leaders	q010210, q010221, q010222, q010223	q010210, q010221, q010222, q010223
Subdomain 1.3: Collaborative design of integrated learning plan	q010330, q010340, q010350^, q010361^, q010362^, q010363^, q010364^, q010365^	q010330, q010340, q010350^, q010361^, q010362^, q010363^, q010364^, q010365^
Subdomain 1.4 Providing appropriate services for students who traditionally struggle	q010420(!), q010421(!), q010422ES(!), q010430T, q010450^, q010461^, q010462^, q010463^, q010464^, q010465^	q010420(!), q010421(!), q010422ES(!), q010430A, q010450^, q010461^, q010462^, q010463^, q010464^, q010465^
Subdomain 2.1 Formative evaluation of student learning	q020110(!), q020120b, q020121(!), q020130, q020130, q020132ES, q020132HS, q020133ES, q020133HS, q020134, q020135	q020110(!), q020120b, q020121(!), q020130, q020130, q020132ES, q020132HS, q020133ES, q020133HS, q020134, q020135
Subdomain 2.2: Summative evaluation of student learning	q020210ES, q020210ES, q020210HS, q020220(!), q020230, q020240, q020241, q020250	q020210ES, q020210ES, q020210HS, q020220(!), q020230, q020240, q020241, q020250
Subdomain 2.3 Formative evaluation of teaching	q020310T_b, q020320, q020334T_b, q020335A_b, q020337T_b, q020338T_b	q020310A_b, q020320, q020334A_b, q020335A_b, q020337A_b, q020338A_b
Subdomain 2.4 Summative evaluation of teaching	q020410(!), q020431T, q020432A, q020433T, q020450T, q020451, q020452, q020460, q020461	q020410(!), q020431A, q020432A, q020433A, q020450A, q020451, q020452, q020460,q020461
Subdomain 3.1 Collaborative school-wide focus on problems of teaching and learning	q030111_b, q030111_b, q030114_b, q030130T_b, q0301301, q030140^, q030150^	q030111_b, q030111_b, q030114_b, q030130A_a, q030130A_b(!), q0301301, q030140^, q030150^
Subdomain 3.2 Professional learning	q030220T, q030230(!), q030240,q030250, q030260^, q030270^, q030280^, q030290^	q030220A, q030230(!), q030240, q030250, q030260^, q030270^, q030280^, q030290^
Subdomain 3.3 Socially distributed leadership	q030311(!), q030312(!), q030313(!), q030325, q030325OU, q030330, q030340	q030311(!), q030312(!), q030313(!), q030325, q030325OU ,q030330, q030340
Subdomain 3.4 Coaching and Mentoring	q030422, q030424, q030432(!), q030440^	q030422, q030424, q030432(!), q030440^
Subdomain 4.1 Personnel	q040113(!), q040115(!),q040117(!),	q040113(!),

practices	q040118ES(!), q040150(!), q040160	q040115(!),q040117(!), q040118ES(!), q040150(!), q040160
Subdomain 4.2 Structuring and maintaining time	q040231(!), q040233(!), q040234(!), q040235ES(!), q040235HS(!), q040236HS(!), q040240^	q040231(!), q040233(!), q040234(!), q040235ES(!), q040235HS(!), q040236HS(!), q040240^
Subdomain 4.3 School resources are focused on student learning	q040313_b(!)^, q040320T, q040330^, q040340, q040350, q040360_b^, q040360_c^	q040313_b(!)^, q040320A, q040330^, q040340, q040350, q040360_b^, q040360_c^
Subdomain 4.4 Integrating external expertise into the school instructional program	040410_b, q040410_d, q040410_d	040410_b, q040410_d, q040410_d
Subdomain 4.5 Coordinating and supervising relations with families and external communities	q040510_c , q040510_d, q040510_e, q040511, q040512, q040520, q040530_b(!), q040560, q040570^, q040580^	q040510_c , q040510_d, q040510_e, q040511, q040512, q040520, q040530_b(!), q040560, q040570^, q040580^
Subdomain 5.1 Clear, consistent, and enforced expectations for student behavior	q050110, q050111_a, q050111_b, q050111_c, q050111_d, q050120, q050121, q050122, q050125, q050130, q050140, q050150, q050160, q050161, q050162, q050163, q050164, q050170	q050110, q050111_a, q050111_b, q050111_c, q050111_d, q050120, q050121, q050122, q050125, q050130, q050140, q050150, q050160, q050161, q050162, q050163, q050164, q050170
Subdomain 5.2 Clean and safe learning environment	q050263, q050267ES, q050268ES, q050271, q050272, q050273,q050277ES, q050278ES	q050250, q050261, q050262, q050263, q050267ES, q050268ES, q050271, q050272, q050273,q050277ES, q050278ES
Subdomain 5.3 Student support services provide safe haven for students who traditionally struggle	q050330_b, q050340, q050351b, q050352b, q050353b, q050354_bHS,q050355b, q050360(!), q050370ES, q050371ES, q050370HS, q050380HS	q050330_b, q050340, q050351b,q050352b, q050353b, q050354_bHS, q050355b, q050360(!), q050370ES, q050371ES, q050370HS, q050380HS
Key: T=teacher A=Administrator ES=Elementary School HS=Middle or High School !=contains “I Don’t Know Responses”		

Appendix C

Descriptive Statistics for CALL Subdomains

	N	Minimum	Maximum	Mean	Std. Deviation
Subdomain 1.1: Maintaining a school-wide focus on learning	8,539	.00	4.00	2.448	.879
Subdomain 1.2 Formal leaders are recognized as instructional leaders	8,501	.00	4.00	2.947	.828
Subdomain 1.3: Collaborative design of integrated learning plan	8,515	.00	4.00	2.685	.830
Subdomain 1.4 Providing appropriate services for students who traditionally struggle	8,435	.00	4.00	2.611	.753
Subdomain 2.1 Formative evaluation of student learning	8,357	.00	4.00	2.539	.801
Subdomain 2.2: Summative evaluation of student learning	8,440	.00	4.00	2.598	.843
Subdomain 2.3 Formative evaluation of teaching	8,269	.00	4.00	1.645	1.037
Subdomain 2.4 Summative evaluation of teaching	8,480	.00	4.00	2.386	.888
Subdomain 3.1 Collaborative school-wide focus on problems of teaching and learning	8,465	.00	4.00	2.490	.850
Subdomain 3.2 Professional learning	8,473	.00	4.00	2.347	.879
Subdomain 3.3 Socially distributed leadership	8,426	.00	4.00	2.347	.926
Subdomain 3.4 Coaching and Mentoring	8,044	.00	4.00	1.511	1.258
Subdomain 4.1 Personnel practices	8,122	.00	4.00	2.657	.972
Subdomain 4.2 Structuring and maintaining time	8,112	.00	4.00	2.516	.878
Subdomain 4.3 School resources are focused on student learning	8,284	.00	4.00	2.109	1.039
Subdomain 4.4 Integrating external expertise into the school instructional program	8,130	.00	4.00	2.007	1.092
Subdomain 4.5 Coordinating and supervising relations with families and external communities	8,146	.00	4.00	2.230	.838
Subdomain 5.1 Clear, consistent, and enforced expectations for student behavior	8,442	.00	4.00	2.623	.742
Subdomain 5.2 Clean and safe learning environment	8,263	.00	4.00	2.512	.732
Subdomain 5.3 Student support services provide safe haven for students who traditionally struggle	8,144	.00	4.00	1.856	.923
Valid N (list wise)	7,931				

Appendix D

Descriptive Statistics for Z-scores used in T-tests and ANOVA Analysis

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness	Kurtosis
Z score: Subdomain 1.1: Maintaining a school-wide focus on learning	8,539	-2.791	1.769	.000	1.00	-.374	-.596
Z score: Subdomain 1.2 Formal leaders are recognized as instructional leaders	8,501	-3.545	1.283	.000	1.00	-.833	.392
Z score: Subdomain 1.3: Collaborative design of integrated learning plan	8,515	-3.233	1.584	.000	1.00	-.414	-.426
Z score: Subdomain 1.4 Providing appropriate services for students who traditionally struggle	8,435	-3.468	1.844	.000	1.00	-.149	-.646
Z score: Subdomain 2.1 Formative evaluation of student learning	8,357	-3.169	1.824	.000	1.00	-.095	-.662
Z score: Subdomain 2.2: Summative evaluation of student learning	8,440	-3.082	1.663	.000	1.00	-.279	-.675
Z score: Subdomain 2.3 Formative evaluation of teaching	8,269	-1.586	2.272	.000	1.00	.427	-.577
Z score: Subdomain 2.4 Summative evaluation of teaching	8,480	-2.686	1.816	.000	1.00	-.191	-.677
Z score: Subdomain 3.1 Collaborative school-wide focus on problems of teaching and learning	8,465	-2.930	1.777	.000	1.00	-.282	-.404
Z score: Subdomain 3.2 Professional learning	8,473	-2.669	1.880	.000	1.00	-.096	-.617
Z score: Subdomain 3.3 Socially distributed leadership	8,426	-2.529	1.791	.000	1.00	-.218	-.777
Z score: Subdomain 3.4 Coaching and Mentoring	8,044	-1.210	1.994	.000	1.00	.383	-1.021
Z score: Subdomain 4.1 Personnel practices	8,122	-2.732	1.381	.000	1.00	-.759	.092
Z score: Subdomain 4.2 Structuring and maintaining time	8,112	-2.865	1.690	.000	1.00	-.430	-.029
Z score: Subdomain 4.3 School resources are focused on student learning	8,284	-2.031	1.821	.000	1.00	.008	-1.002

Z score: Subdomain 4.4 Integrating external expertise into the school instructional program	8,130	-1.839	1.825	.000	1.00	-.147	-.910
Z score: Subdomain 4.5 Coordinating and supervising relations with families and external communities	8,146	-2.662	2.112	.000	1.00	-.201	-.625
Z score: Subdomain 5.1 Clear, consistent, and enforced expectations for student behavior	8,442	-3.534	1.855	.000	1.00	-.397	-.091
Z score: Subdomain 5.2 Clean and safe learning environment	8,263	-3.430	2.031	.000	1.00	-.371	.484
Z score: Subdomain 5.3 Student support services provide safe haven for students who traditionally struggle	8,144	-2.011	2.322	.000	1.00	.112	-.751
Valid N (listwise)	7,931						

Appendix E

Descriptive Statistics for Z-scores Separated by Teacher and Administrator Groups

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness	Kurtosis
TEACHERS							
Z score: Subdomain 1.1: Maintaining a school-wide focus on learning	8,005	-2.791	1.769	-.027	1.003	-.347	-.625
Z score: Subdomain 1.2 Formal leaders are recognized as instructional leaders	7,975	-3.555	1.283	-.024	1.002	-.816	.366
Z score: Subdomain 1.3: Collaborative design of integrated learning plan	7,988	-3.223	1.584	-.009	1.002	-.407	-.440
Z score: Subdomain 1.4 Providing appropriate services for students who traditionally struggle	7,919	-3.468	1.844	-.005	1.000	-.144	-.654
Z score: Subdomain 2.1 Formative evaluation of student learning	7,852	-3.169	1.824	-.012	.999	-.082	-.670
Z score: Subdomain 2.2: Summative evaluation of student learning	7,922	-3.082	1.663	-.011	1.004	-.277	-.683
Z score: Subdomain 2.3 Formative evaluation of teaching	7,769	-1.586	2.272	-.022	.998	.452	-.553
Z score: Subdomain 2.4 Summative evaluation of teaching	7,959	-2.686	1.817	-.023	1.003	-.167	-.692
Z score: Subdomain 3.1 Collaborative school-wide focus on problems of teaching & learning	7,947	-2.931	1.777	-.017	1.001	-.269	-.407
Z score: Subdomain 3.2 Professional learning	7,954	-2.669	1.880	-.017	1.001	-.078	-.619
Z score: Subdomain 3.3 Socially distributed leadership	7,906	-2.529	1.791	-.015	1.002	-.201	-.790
Z score: Subdomain 3.4 Coaching and Mentoring	7,555	-1.210	1.994	-.011	.999	.402	-1.002
Z score: Subdomain 4.1 Personnel practices	7,634	-2.732	1.381	-.020	1.003	-.730	.038
Z score: Subdomain 4.2 Structuring and maintaining time	7,626	-2.865	1.690	-.013	1.005	-.409	-.059
Z score: Subdomain 4.3 School resources are focused on student learning	7,779	-2.031	1.821	-.021	1.001	.037	-.998
Z score: Subdomain 4.4 Integrating external expertise into the school instructional program	7,639	-1.839	1.825	-.014	.999	-.132	-.913
Z score: Subdomain 4.5 Coordinating & supervising relations with families & external communities	7,656	-2.662	2.112	-.019	1.002	-.183	-.637
Z score: Subdomain 5.1 Clear, consistent, and enforced expectations for student behavior	7,925	-3.534	1.855	-.021	1.002	-.387	-.099
Z score: Subdomain 5.2 Clean and safe learning environment	7,764	-3.430	2.031	-.009	1.001	-.353	.457
Z score: Subdomain 5.3 Student support services provide safe haven for students who traditionally struggle	7,654	-2.011	2.323	-.015	1.002	.136	-.745
Valid N (listwise)	7,453						

ADMINISTRATORS

Z score: Subdomain 1.1: Maintaining a school-wide focus on learning	534	-2.791	1.769	.407	.853	-.742	.305
Z score: Subdomain 1.2 Formal leaders are recognized as instructional leaders	526	-2.941	1.283	.367	.882	-1.145	1.164
Z score: Subdomain 1.3: Collaborative design of integrated learning plan	527	-2.932	1.584	.139	.956	-.508	-.161
Z score: Subdomain 1.4 Providing appropriate services for students who traditionally struggle	516	-3.468	1.844	.079	1.005	-.231	-.503
Z score: Subdomain 2.1 Formative evaluation of student learning	505	-3.169	1.824	.190	.999	-.308	-.417
Z score: Subdomain 2.2: Summative evaluation of student learning	518	-2.489	1.664	.174	.928	-.245	-.671
Z score: Subdomain 2.3 Formative evaluation of teaching	500	-1.586	2.272	.349	.962	.128	-.596
Z score: Subdomain 2.4 Summative evaluation of teaching	521	-2.686	1.817	.344	.882	-.512	-.121
Z score: Subdomain 3.1 Collaborative school-wide focus on problems of teaching and learning	518	-2.931	1.777	.258	.957	-.482	.-222
Z score: Subdomain 3.2 Professional learning	519	-2.669	1.880	.255	.949	-.350	-.387
Z score: Subdomain 3.3 Socially distributed leadership	520	-2.529	1.791	.225	.943	-.464	-.409
Z score: Subdomain 3.4 Coaching and Mentoring	489	-1.210	1.994	.171	.998	.095	-1.165
Z score: Subdomain 4.1 Personnel practices	488	-2.732	1.381	.311	.893	-1.321	1.852
Z score: Subdomain 4.2 Structuring and maintaining time	486	-2.865	1.690	.199	.894	-.756	.788
Z score: Subdomain 4.3 School resources are focused on student learning	505	-2.031	1.821	.331	.931	-.426	-.675
Z score: Subdomain 4.4 Integrating external expertise into the school instructional program	491	-1.839	1.825	.208	.992	-.405	-.716
Z score: Subdomain 4.5 Coordinating and supervising relations with families & external communities	490	-2.662	2.112	.304	.914	-.440	-.224
Z score: Subdomain 5.1 Clear, consistent, and enforced expectations for student behavior	517	-2.996	1.855	.318	.915	-.497	.051
Z score: Subdomain 5.2 Clean and safe learning environment	499	-3.430	2.031	.144	.979	-.654	1.160
Z score: Subdomain 5.3 Student support services provide safe haven for students who traditionally struggle	490	-2.011	2.323	.228	.946	-.248	-.562
Valid N (listwise)	478						

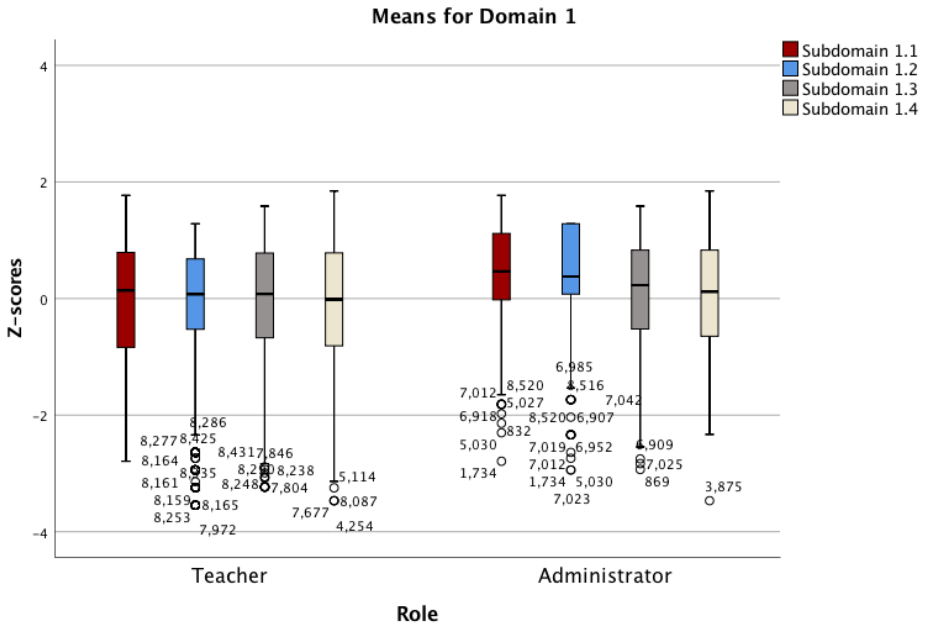
Appendix F

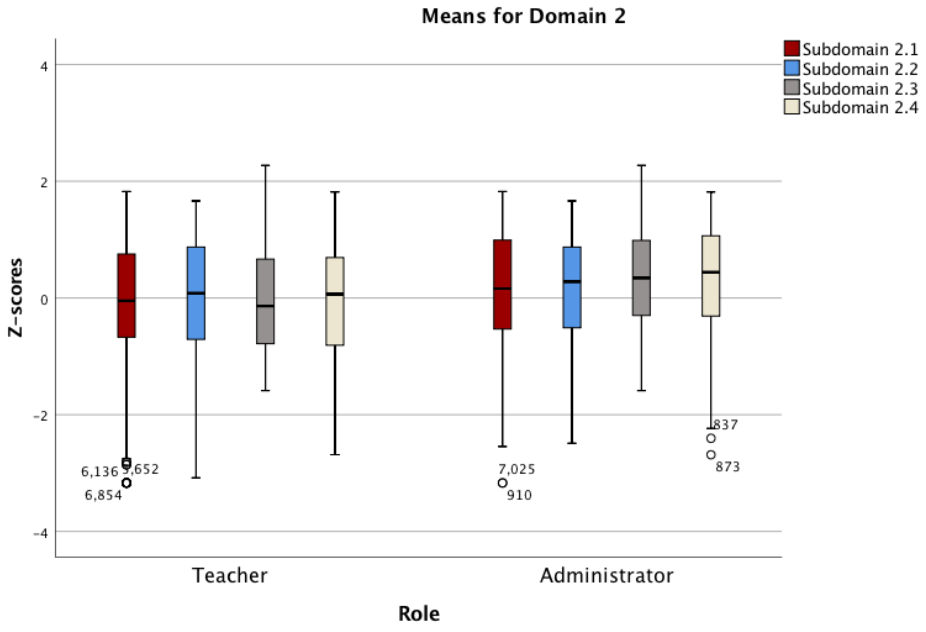
Degree of Discrepancy Between Teachers and Leaders: Quartile List of Subdomains
Appendix G

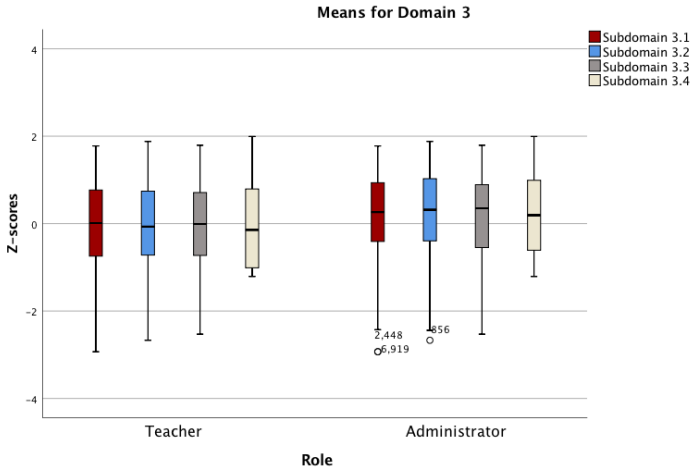
			95% Confidence Interval of the Difference		
	Mean Difference	Std. Error Difference	Lower	Upper	Quartile
Z score: Subdomain 1.1: Maintaining a school-wide focus on learning	-.435	.039	-.510	-.359	4
Z score: Subdomain 1.2 Formal leaders are recognized as instructional leaders	-.392	.040	-.470	-.313	4
Z score: Subdomain 2.3 Formative evaluation of teaching	-.371	.044	-.459	-.284	4
Z score: Subdomain 2.4 Summative evaluation of teaching	-.367	.040	-.446	-.288	4
Z score: Subdomain 4.3 School resources are focused on student learning	-.353	.043	-.437	-.268	4
Z score: Subdomain 5.1 Clear, consistent, and enforced expectations for student behavior	-.338	.042	-.420	-.256	4
Z score: Subdomain 4.1 Personnel practices	-.331	.042	-.413	-.248	3
Z score: Subdomain 4.5 Coordinating and supervising relations with families and external communities	-.324	.043	-.408	-.239	3
Z score: Subdomain 3.1 Collaborative school-wide focus on problems of teaching and learning	-.275	.044	-.361	-.190	3
Z score: Subdomain 3.2 Professional learning	-.272	.043	-.357	-.187	3
Z score: Subdomain 5.3 Student support services provide safe haven for students who traditionally struggle	-.242	.044	-.329	-.155	2
Z score: Subdomain 3.3 Socially distributed leadership	-.240	.043	-.325	-.156	2
Z score: Subdomain 4.4 Integrating external expertise into the school instructional program	-.222	.046	-.313	-.131	2
Z score: Subdomain 4.2 Structuring and maintaining time	-.212	.042	-.295	-.129	2
Z score: Subdomain 2.1 Formative evaluation of student learning	-.203	.046	-.293	-.113	1
Z score: Subdomain 2.2: Summative evaluation of student learning	-.185	.042	-.268	-.102	1
Z score: Subdomain 3.4 Coaching and Mentoring	-.183	.047	-.274	-.091	1
Z score: Subdomain 5.2 Clean and safe learning environment	-.153	.045	-.242	-.065	1
Z score: Subdomain 1.3: Collaborative design of integrated learning plan	-.148	.043	-.233	-.063	1
Z score: Subdomain 1.4 Providing appropriate services for students who traditionally struggle	-.084	.046	-.173	.006	1

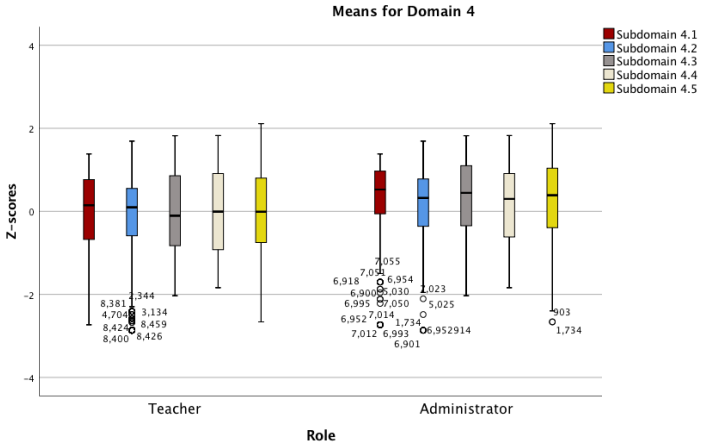
Appendix G

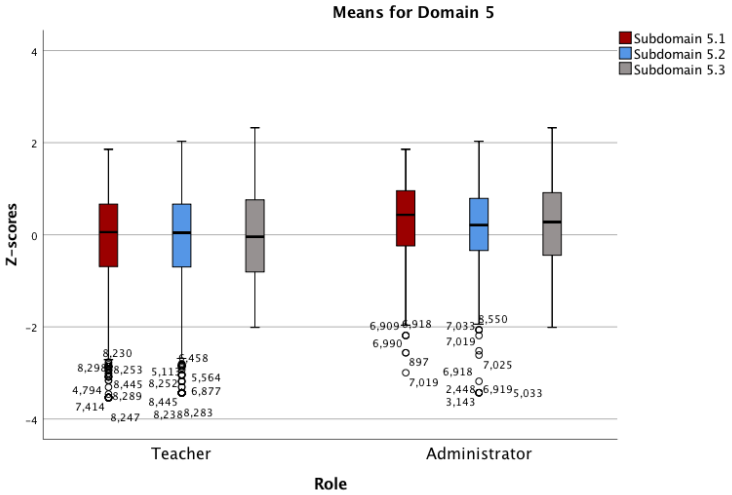
Box and Whisker Plots of the Difference in Means for Teachers and Administrators by Subdomain






Means for Domain 3
Subdomain 3.1
Subdomain 3.2
Subdomain 3.3
Subdomain 3.4
Z-scores
4
2
0
-2
-4
2,448
6,919
856
Teacher
Administrator
Role


Means for Domain 4
Subdomain 4.1
Subdomain 4.2
Subdomain 4.3
Subdomain 4.4
Subdomain 4.5
Z-scores
4
2
0
-2
-4
8,381
4,704
8,424
8,400
8,426
8,459
3,134
2,344
7,055
7,054
6,918
6,954
6,900
5,030
6,995
7,050
5,025
1,023
6,952
7,014
1,734
7,012
6,993
6,901
6,952914
903
1,734
Teacher
Administrator
Role


Means for Domain 5
Subdomain 5.1
Subdomain 5.2
Subdomain 5.3
Z-scores
4
2
0
-2
-4
8,230
8,253
8,445
8,289
8,247
4,794
7,414
8,445
8,238
8,283
5,564
6,877
6,909
4,918
6,990
897
7,019
7,033
3,550
7,019
6,918
7,025
2,448
3,143
6,919
5,033
Teacher
Administrator
Role

Appendix H

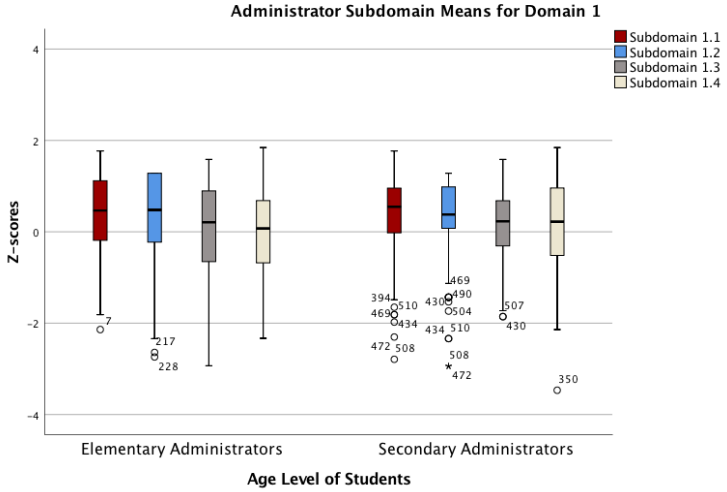
Descriptive Statistics for Independent Samples T-test between Elementary and Secondary Administrators

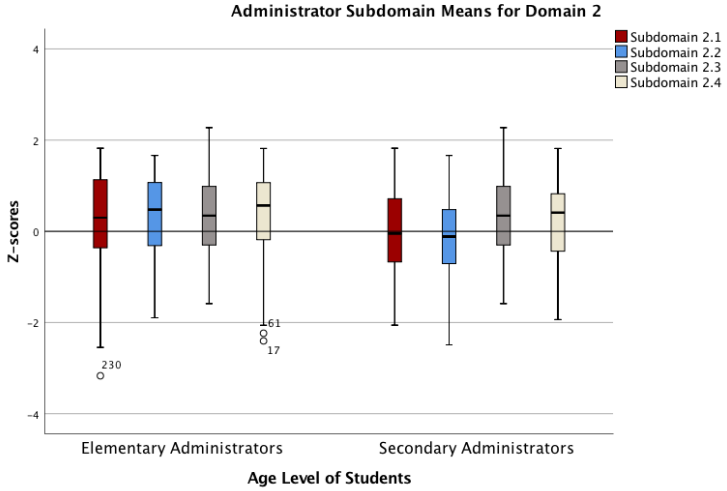
	ES or MSHS	N	Mean	Min	Max	Std. Deviation	Std. Error Mean
Z score: Subdomain 1.1: Maintaining a school-wide focus on learning	Elementary	333	.407	-2.140	1.769	.871	.048
	Secondary	201	.408	-2.791	1.769	.824	.058
Z score: Subdomain 1.2 Formal leaders are recognized as instructional leaders	Elementary	326	.363	-2.740	1.283	.900	.050
	Secondary	200	.375	-2.941	1.283	.854	.060
Z score: Subdomain 1.3: Collaborative design of integrated learning plan	Elementary	328	.118	-2.932	1.584	1.027	.057
	Secondary	199	.173	-2.029	1.584	.829	.059
Z score: Subdomain 1.4 Providing appropriate services for students who traditionally struggle	Elementary	321	.002	-2.330	1.844	.995	.056
	Secondary	195	.205	-3.468	1.844	1.010	.072
Z score: Subdomain 2.1 Formative evaluation of student learning	Elementary	314	.291	-3.169	1.824	1.002	.057
	Secondary	191	.024	-2.059	1.824	.973	.070
Z score: Subdomain 2.2: Summative evaluation of student learning	Elementary	321	.331	-1.896	1.664	.874	.049
	Secondary	197	-.083	-2.489	1.664	.958	.068
Z score: Subdomain 2.3 Formative evaluation of teaching	Elementary	312	.347	-1.586	2.272	1.002	.057
	Secondary	188	.353	-1.586	2.272	.893	.065
Z score: Subdomain 2.4 Summative evaluation of teaching	Elementary	324	.384	-2.686	1.817	.908	.050
	Secondary	197	.279	-1.936	1.817	.835	.060
Z score: Subdomain 3.1 Collaborative school-wide focus on problems of teaching and learning	Elementary	322	.281	-2.931	1.777	.990	.055
	Secondary	196	.222	-1.922	1.777	.900	.064
Z score: Subdomain 3.2 Professional learning	Elementary	323	.241	-2.669	1.880	.944	.053
	Secondary	196	.280	-2.441	1.880	.960	.069
Z score: Subdomain 3.3 Socially distributed leadership	Elementary	323	.176	-2.529	1.791	.966	.054
	Secondary	197	.306	-2.529	1.791	.901	.064
Z score: Subdomain 3.4 Coaching and Mentoring	Elementary	303	.195	-1.210	1.994	1.024	.059
	Secondary	186	.133	-1.210	1.994	.958	.070
Z score: Subdomain 4.1 Personnel practices	Elementary	301	.294	-2.732	1.381	.985	.057
	Secondary	187	.338	-2.732	1.381	.723	.053
Z score: Subdomain 4.2	Elementary	300	.249	-2.865	1.690	.972	.056

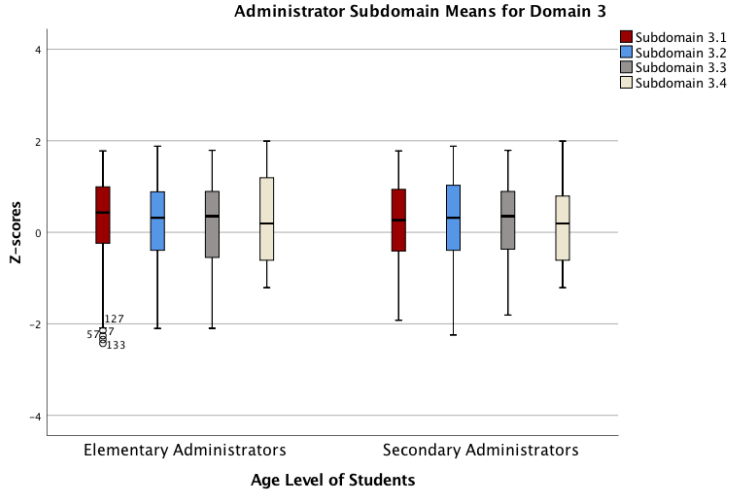
Structuring and maintaining time	Secondary	186	.118	-2.865	1.690	.747	.055
Z score: Subdomain 4.3	Elementary	319	.311	-2.031	1.821	.912	.051
School resources are focused on student learning	Secondary	186	.366	-2.031	1.821	.964	.071
Z score: Subdomain 4.4	Elementary	304	.126	-1.839	1.825	1.005	.058
Integrating external expertise into the school instructional program	Secondary	187	.343	-1.839	1.825	.958	.070
Z score: Subdomain 4.5	Elementary	303	.265	-2.662	2.112	.943	.054
Coordinating and supervising relations with families and external communities	Secondary	187	.367	-2.662	1.874	.863	.063
Z score: Subdomain 5.1	Elementary	320	.284	-2.996	1.855	.953	.053
Clear, consistent, and enforced expectations for student behavior	Secondary	197	.371	-1.963	1.855	.849	.060
Z score: Subdomain 5.2	Elementary	309	.078	-3.430	2.031	1.001	.057
Clean and safe learning environment	Secondary	190	.252	-3.430	2.031	.936	.068
Z score: Subdomain 5.3	Elementary	304	.166	-2.011	2.323	.983	.056
Student support services provide safe haven for students who traditionally struggle	Secondary	186	.329	-1.686	2.323	.877	.064

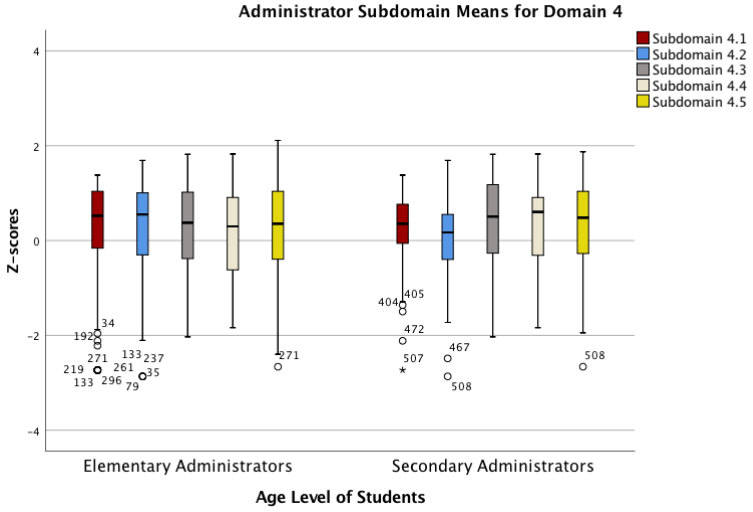
Appendix I

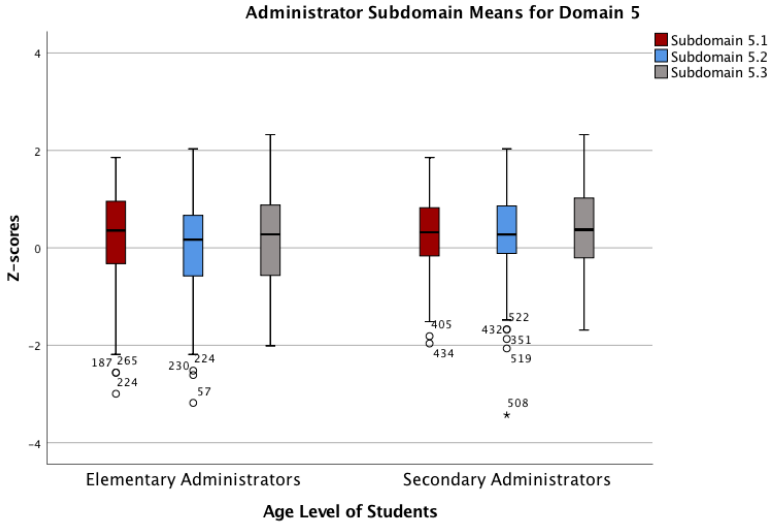
Box and Whisker Plots of the Difference in Means for Secondary and Elementary Administrators by Domain






Administrator Subdomain Means for Domain 3
Subdomain 3.1
Subdomain 3.2
Subdomain 3.3
Subdomain 3.4
Z-scores
4
2
0
-2
-4
127
57
133
Elementary Administrators
Secondary Administrators
Age Level of Students


Administrator Subdomain Means for Domain 4
Subdomain 4.1
Subdomain 4.2
Subdomain 4.3
Subdomain 4.4
Subdomain 4.5
Z-scores
4
2
0
-2
-4
34
192
271
219
133
261
296
133
237
35
79
271
404
405
472
507
467
508
508
Elementary Administrators
Secondary Administrators
Age Level of Students


Administrator Subdomain Means for Domain 5
Subdomain 5.1
Subdomain 5.2
Subdomain 5.3
Z-scores
4
2
0
-2
-4
187
265
224
230
224
57
405
434
432
522
351
519
508
Elementary Administrators
Secondary Administrators
Age Level of Students

Appendix J

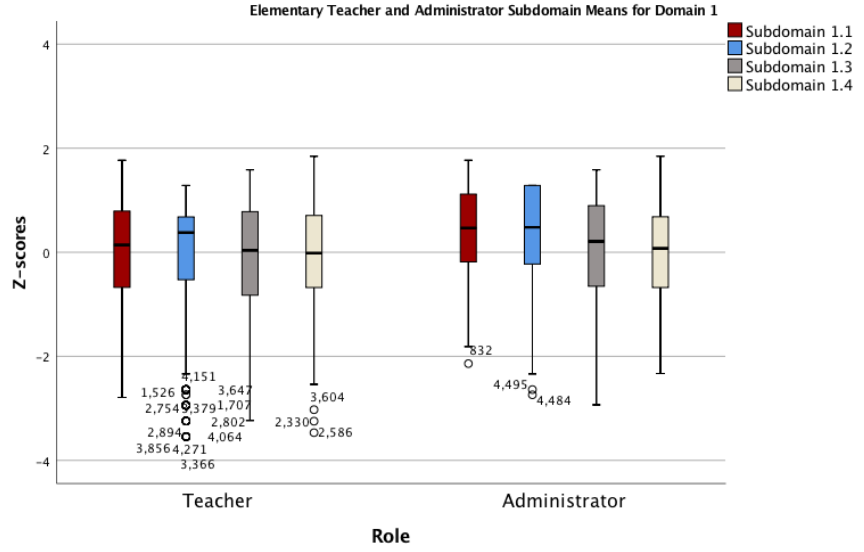
Degree of Discrepancy Between Teachers and Leaders: Quartiles of Subdomains by School Level

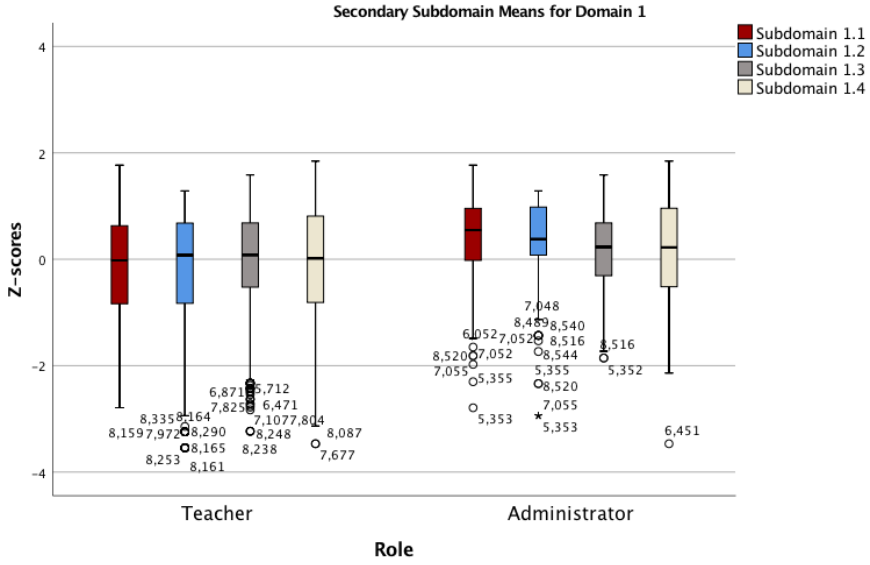
			95% Confidence Interval of the Difference		
	Mean Difference	SE Difference	Lower	Upper	Quartile
ELEMENTARY SCHOOLS					
Z score: Subdomain 1.1: Maintaining a school-wide focus on learning	-.333	.050	-.431	-.234	4
Z score: Subdomain 2.4 Summative evaluation of teaching	-.313	.053	-.417	-.209	4
Z score: Subdomain 2.3 Formative evaluation of teaching	-.304	.059	-.419	-.188	4
Z score: Subdomain 4.1 Personnel practices	-.294	.059	-.410	-.178	4
Z score: Subdomain 1.2 Formal leaders are recognized as instructional leaders	-.284	.052	-.387	-.182	4
Z score: Subdomain 4.3 School resources are focused on student learning	-.269	.053	-.373	-.164	4
Z score: Subdomain 2.2: Summative evaluation of student learning	-.193	.051	-.293	-.093	3
Z score: Subdomain 4.5 Coordinating and supervising relations with families and external communities	-.186	.056	-.297	-.075	3
Z score: Subdomain 1.3: Collaborative design of integrated learning plan	-.180	.059	-.296	-.064	3
Z score: Subdomain 3.2 Professional learning	-.179	.055	-.287	-.072	3
Z score: Subdomain 2.1 Formative evaluation of student learning	-.169	.059	-.284	-.054	2
Z score: Subdomain 3.3 Socially distributed leadership	-.166	.056	-.276	-.056	2
Z score: Subdomain 5.1 Clear, consistent, and enforced expectations for student behavior	-.165	.055	-.274	-.056	2
Z score: Subdomain 3.1 Collaborative school-wide focus on problems of teaching and learning	-.158	.057	-.270	-.045	2
Z score: Subdomain 4.2 Structuring and maintaining time	-.140	.058	-.255	-.025	1
Z score: Subdomain 5.3 Student support services provide safe haven for students who traditionally struggle	-.123	.059	-.238	-.008	1
Z score: Subdomain 3.4 Coaching and Mentoring	-.096	.061	-.216	.023	1
Z score: Subdomain 4.4 Integrating external expertise into the school instructional program	-.074	.060	-.191	.044	1
Z score: Subdomain 5.2 Clean and safe	-.047	.059	-.163	.069	1

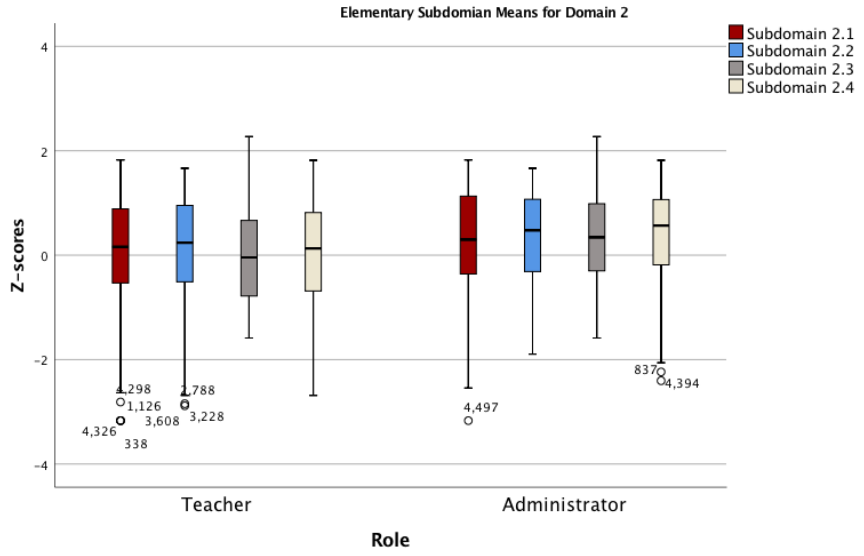
learning environment					
Z score: Subdomain 1.4 Providing appropriate services for students who traditionally struggle	-.002	.057	-.115	.111	1
SECONDARY SCHOOLS					
Z score: Subdomain 5.1 Clear, consistent, and enforced expectations for student behavior	-.547	.063	-.671	-.424	4
Z score: Subdomain 1.1: Maintaining a school-wide focus on learning	-.547	.060	-.666	-.428	4
Z score: Subdomain 1.2 Formal leaders are recognized as instructional leaders	-.513	.063	-.636	-.389	4
Z score: Subdomain 4.5 Coordinating and supervising relations with families and external communities	-.496	.065	-.625	-.368	4
Z score: Subdomain 4.3 School resources are focused on student learning	-.461	.073	-.604	-.317	4
Z score: Subdomain 2.3 Formative evaluation of teaching	-.447	.067	-.580	-.315	4
Z score: Subdomain 4.4 Integrating external expertise into the school instructional program	-.429	.072	-.571	-.287	3
Z score: Subdomain 5.3 Student support services provide safe haven for students who traditionally struggle	-.407	.066	-.538	-.276	3
Z score: Subdomain 2.4 Summative evaluation of teaching	-.405	.062	-.526	-.283	3
Z score: Subdomain 3.1 Collaborative school-wide focus on problems of teaching and learning	-.393	.066	-.523	-.262	3
Z score: Subdomain 3.2 Professional learning	-.382	.070	-.521	-.244	2
Z score: Subdomain 4.1 Personnel practices	-.379	.055	-.488	-.270	2
Z score: Subdomain 3.3 Socially distributed leadership	-.349	.066	-.480	-.219	2
Z score: Subdomain 5.2 Clean and safe learning environment	-.306	.070	-.444	-.168	2
Z score: Subdomain 3.4 Coaching and Mentoring	-.266	.072	-.409	-.124	1
Z score: Subdomain 4.2 Structuring and maintaining time	-.266	.057	-.378	-.154	1
Z score: Subdomain 1.4 Providing appropriate services for students who traditionally struggle	-.216	.074	-.362	-.069	1
Z score: Subdomain 2.1 Formative evaluation of student learning	-.186	.072	-.328	-.043	1
Z score: Subdomain 1.3: Collaborative design of integrated learning plan	-.124	.061	-.243	-.004	1

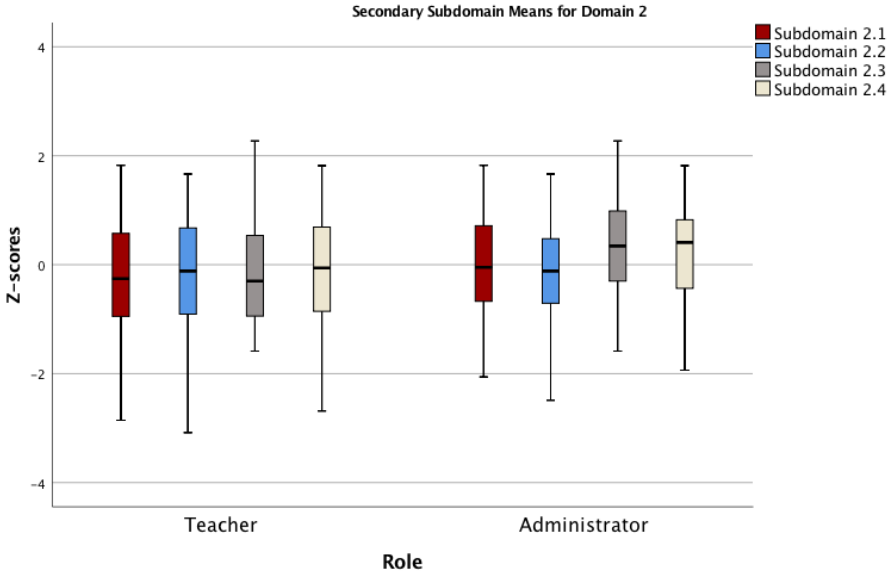
Appendix K

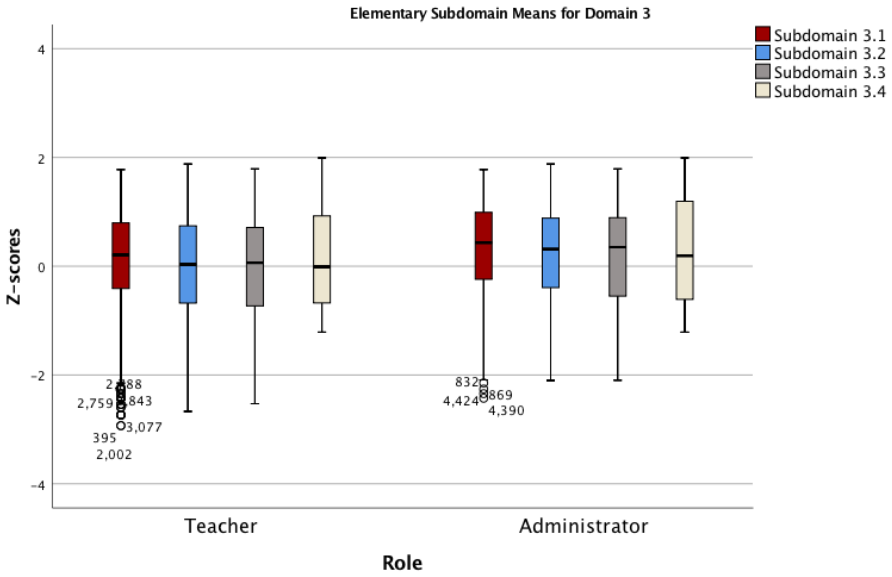
Box and Whisker Plots of the Difference in Teacher and Administrator Subdomain Means by School Level

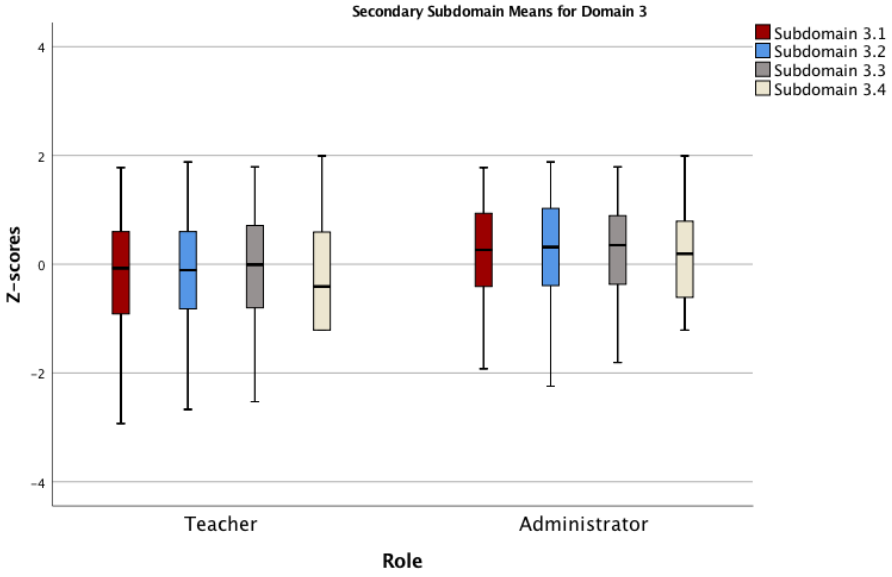


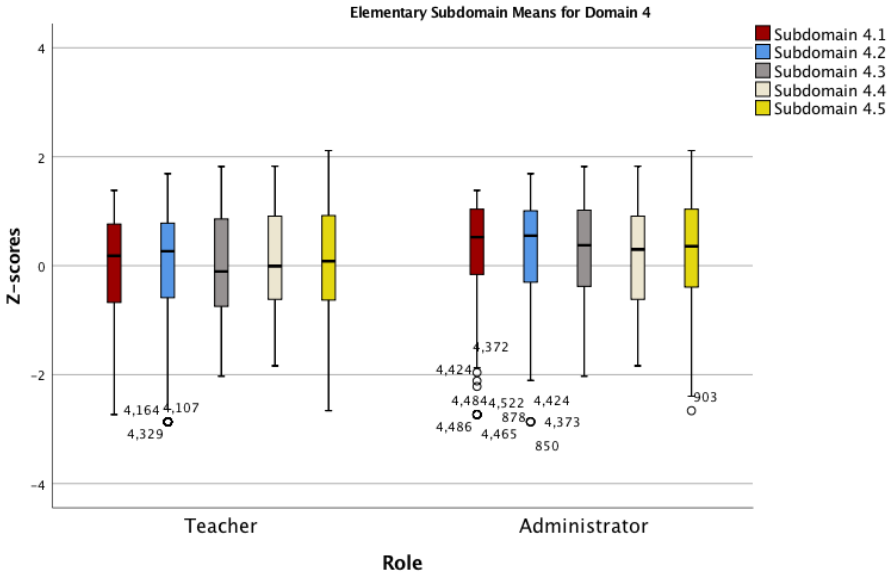


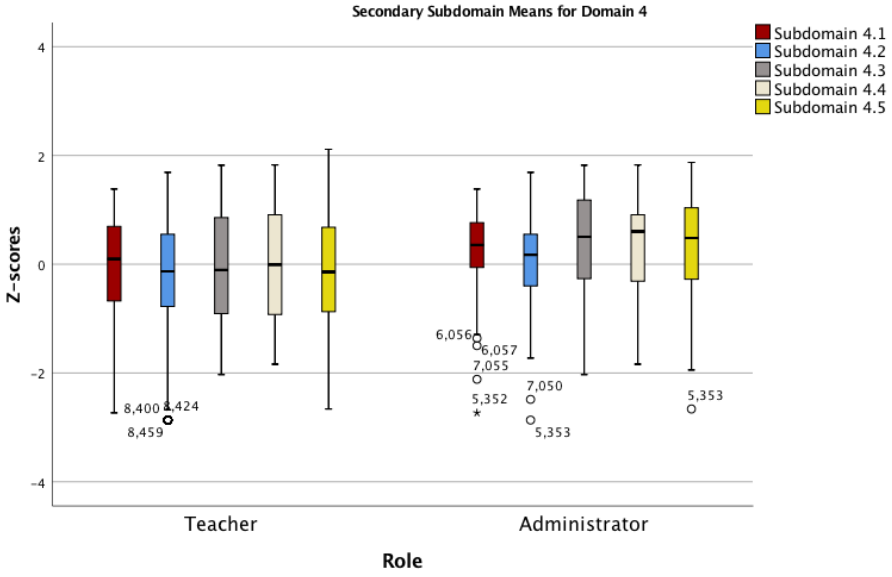

Elementary Subdomian Means for Domain 2
Subdomain 2.1
Subdomain 2.2
Subdomain 2.3
Subdomain 2.4
Z-scores
4
2
0
-2
-4
4,298
1,126
4,326
3,608
338
1,788
3,228
4,497
837
4,394
Teacher
Administrator
Role

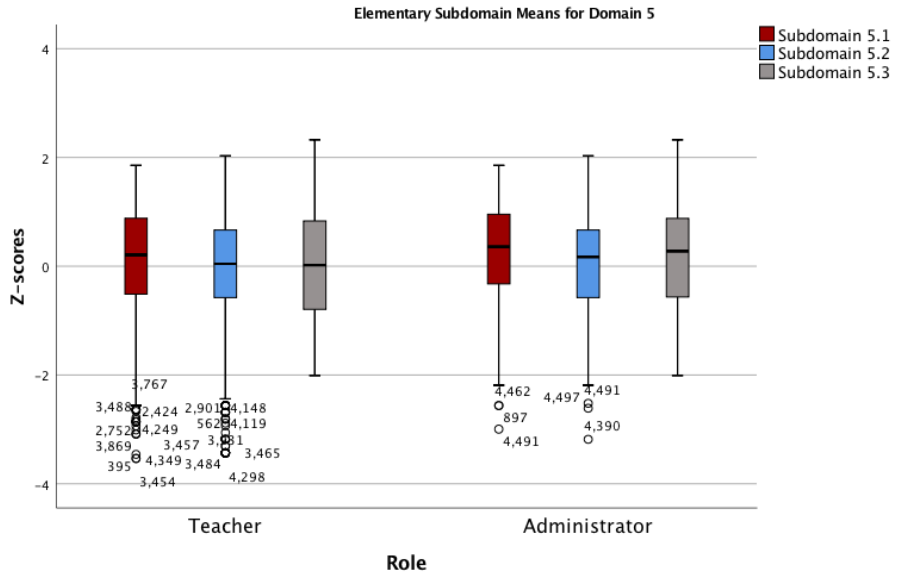

Secondary Subdomain Means for Domain 2
Subdomain 2.1
Subdomain 2.2
Subdomain 2.3
Subdomain 2.4
Z-scores
4
2
0
-2
-4
Teacher
Administrator
Role

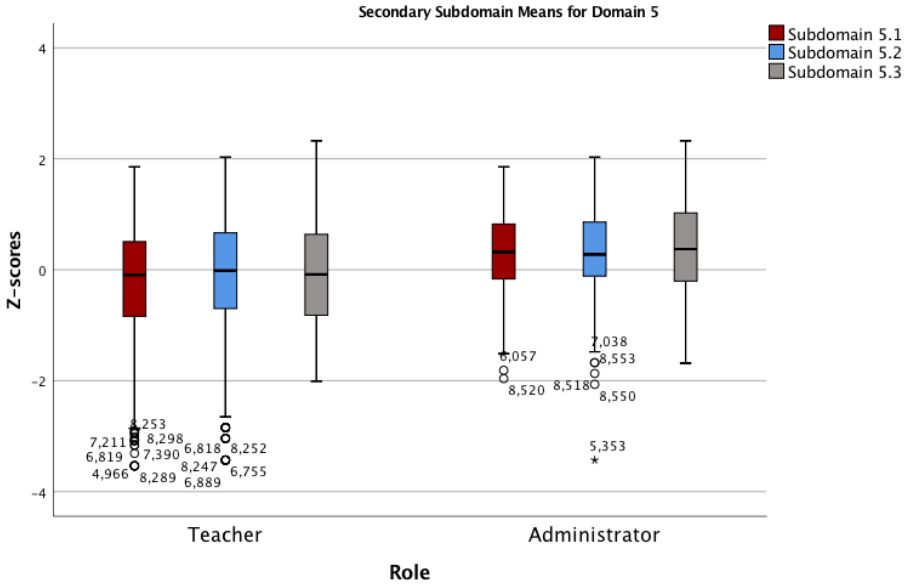

Elementary Subdomain Means for Domain 3
Subdomain 3.1
Subdomain 3.2
Subdomain 3.3
Subdomain 3.4
Z-scores
4
2
0
-2
-4
2,759
3,077
395
2,002
832
4,424
4,390
Teacher
Administrator
Role


Secondary Subdomain Means for Domain 3
Subdomain 3.1
Subdomain 3.2
Subdomain 3.3
Subdomain 3.4
Z-scores
4
2
0
-2
-4
Teacher
Administrator
Role


Elementary Subdomain Means for Domain 4
Subdomain 4.1
Subdomain 4.2
Subdomain 4.3
Subdomain 4.4
Subdomain 4.5
Z-scores
4
2
0
-2
-4
4,164
4,107
4,329
4,372
4,424
4,484
4,522
4,424
878
4,373
4,486
4,465
850
903
Teacher
Administrator
Role


Secondary Subdomain Means for Domain 4
Subdomain 4.1
Subdomain 4.2
Subdomain 4.3
Subdomain 4.4
Subdomain 4.5
Z-scores
4
2
0
-2
-4
8,400
8,424
8,459
6,056
6,057
7,055
5,352
7,050
5,353
5,353
Teacher
Administrator
Role


Elementary Subdomain Means for Domain 5
Subdomain 5.1
Subdomain 5.2
Subdomain 5.3
Z-scores
4
2
0
-2
-4
Teacher
Administrator
Role


Secondary Subdomain Means for Domain 5
Subdomain 5.1
Subdomain 5.2
Subdomain 5.3
Z-scores
4
2
0
-2
-4
Teacher
Administrator
Role

Appendix L

Descriptive Statistics for ANOVA Test

			N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean Lower Bound	95% Confidence Interval for Mean Upper Bound	Min	Max	Between-Component Variance
Z score: Subdomain 1.1: Maintaining a school-wide focus on learning		Rural	73	.306	.779	.091	.125	.488	-1.651	1.443	
		Small Town	28	.644	.529	.100	.438	.849	-.511	1.443	
		Suburban	46	.413	.865	.128	.156	.670	-2.303	1.769	
		Urban	267	.425	.839	.051	.324	.526	-1.814	1.769	
		Total	414	.417	.815	.040	.339	.496	-2.303	1.769	
		Fixed Effects		.815	.040	.339	.496				
	Model	Random Effects			.048	.263	.572				.002
Z score: Subdomain 1.2 Formal leaders are recognized as instructional leaders		Rural	72	.546	.662	.078	.390	.701	-1.433	1.283	
		Small Town	28	.464	.512	.097	.266	.663	-1.131	1.283	
		Suburban	46	.404	.855	.126	.150	.658	-2.941	1.283	
		Urban	261	.292	.942	.058	.177	.407	-2.740	1.283	
		Total	407	.361	.868	.043	.277	.446	-2.941	1.283	
		Fixed Effects		.866	.043	.277	.446				
	Model	Random Effects			.075	.122	.601				.008
Z score: Subdomain 1.3: Collaborative design of integrated learning plan		Rural	72	.029	.838	.099	-.167	.226	-2.229	1.584	
		Small Town	28	.403	.759	.144	.108	.697	-2.029	1.584	
		Suburban	45	.153	.686	.102	-.053	.359	-1.306	1.434	
		Urban	263	.133	.978	.060	.014	.252	-2.832	1.584	
		Total	408	.136	.913	.045	.047	.224	-2.832	1.584	
		Fixed Effects		.913	.045	.047	.224				
	Model	Random Effects			.052	-.031	.302				.001
Z score: Subdomain 1.4 Providing appropriate services for students who traditionally struggle		Rural	69	.129	.931	.112	-.095	.352	-2.007	1.844	
		Small Town	27	.375	1.128	.217	-.071	.821	-3.468	1.844	
		Suburban	45	.533	.733	.109	.313	.754	-.960	1.844	
		Urban	259	-.043	1.041	.065	-.171	.084	-2.140	1.844	
		Total	400	.080	1.015	.051	-.020	.179	-3.468	1.844	
		Fixed Effects		.999	.050	-.019	.178				
	Model	Random Effects			.174	-.474	.633				.059
Z score: Subdomain 2.1 Formative evaluation of student learning		Rural	68	.044	.917	.111	-.178	.266	-1.643	1.824	
		Small Town	27	.241	.815	.157	-.082	.563	-1.504	1.408	
		Suburban	44	-.169	.954	.144	-.459	.121	-2.059	1.824	
		Urban	252	.246	1.012	.064	.120	.371	-3.169	1.824	
		Total	391	.164	.984	.050	.066	.261	-3.169	1.824	
		Fixed Effects		.978	.049	.066	.261				
	Model	Random Effects			.114	-.201	.528				.023
Z score: Subdomain		Rural	70	-.063	.860	.103	-.268	.142	-1.896	1.664	
		Small Town	27	.076	.847	.163	-.259	.411	-1.698	1.466	

2.2: Summative evaluation of student learning		Suburban	46	-.225	.974	.144	-.515	.064	-2.370	1.664	
		Urban	259	.306	.892	.055	.197	.415	-2.094	1.664	
		Total	402	.166	.912	.046	.076	.255	-2.370	1.664	
		Fixed Effects		.894	.045	.078	.253				
	Model	Random Effects			.177	-.397	.728				.063
		Rural	68	.365	.767	.093	.180	.551	-1.426	2.272	
		Small Town	26	.561	.772	.151	.249	.872	-1.265	2.272	
		Suburban	42	.213	.800	.123	-.037	.462	-1.265	1.790	
Z score: Subdomain 2.3 Formative evaluation of teaching		Urban	251	.308	1.019	.064	.181	.435	-1.586	2.272	
		Total	387	.325	.942	.048	.231	.419	-1.586	2.272	
		Fixed Effects		.942	.048	.231	.419				
	Model	Random Effects			.048[a]	172[a]	.477[a]				-.002
		Rural	71	.270	.752	.089	.092	.448	-1.560	1.817	
		Small Town	28	.339	.583	.110	.113	.565	-.575	1.366	
		Suburban	45	.046	.711	.106	-.168	.260	-1.560	1.692	
Z score: Subdomain 2.4 Summative evaluation of teaching		Urban	260	.413	.902	.056	.302	.523	-2.236	1.817	
		Total	404	.342	.844	.042	.259	.424	-2.236	1.817	
		Fixed Effects		.839	.042	.259	.424				
	Model	Random Effects			.096	.037	.647				.016
		Rural	70	.179	.896	.107	-.035	.392	-2.931	1.777	
Z score: Subdomain 3.1 Collaborative school-wide focus on problems of teaching and learning		Small Town	27	.428	.772	.149	.123	.734	-.969	1.441	
		Suburban	45	.209	.885	.132	-.056	.475	-1.754	1.777	
		Urban	260	.260	1.000	.062	.138	.382	-2.931	1.777	
		Total	402	.252	.955	.048	.158	.345	-2.931	1.777	
		Fixed Effects		.957	.048	.158	.345				
	Model	Random Effects			.048[a]	.100[a]	403[a]				-.007
		Rural	70	.174	.939	.112	-.050	.398	-1.958	1.880	
		Small Town	27	.424	.792	.152	.111	.737	-1.105	1.880	
		Suburban	45	.117	.851	.127	-.138	.373	-2.242	1.880	
		Urban	260	.258	.944	.059	.143	.374	-2.100	1.880	
Z score: Subdomain 3.2 Professional learning		Total	402	.239	.923	.046	.148	.329	-2.242	1.880	
		Fixed Effects		.924	.046	.148	.330				
	Model	Random Effects			.046[a]	092[a]	.386[a]				-.003
		Rural	70	.140	.897	.107	-.074	.354	-1.665	1.791	
		Small Town	27	.494	.768	.148	.190	.798	-1.269	1.791	
		Suburban	46	.127	.840	.124	-.122	.377	-1.269	1.791	
		Urban	260	.215	.912	.057	.104	.326	-2.529	1.791	
Z score: Subdomain 3.3 Socially distributed leadership		Total	403	.211	.893	.045	.123	.298	-2.529	1.791	
		Fixed Effects		.893	.044	.123	.298				
	Model	Random Effects			.054	.038	.384				.002
		Rural	66	.165	.887	.109	-.053	.383	-1.210	1.793	
Z score: Subdomain 3.4 Coaching and Mentoring		Small Town	26	.327	.916	.180	-.043	.697	-1.210	1.994	
		Suburban	42	.150	.926	.143	-.139	.439	-1.210	1.994	
		Urban	246	.285	1.023	.065	.156	.413	-1.210	1.994	
		Total	380	.252	.982	.050	.153	.351	-1.210	1.994	

		Fixed Effects		.984	.050	.153	.351				
	Model	Random Effects			.050[a]	.091[a]	.412[a]				-.008
		Rural	66	.282	.736	.091	.101	.463	-1.498	1.381	
		Small Town	26	.581	.442	.087	.402	.760	-.161	1.381	
		Suburban	42	.525	.818	.126	.270	.780	-2.115	1.381	
		Urban	245	.275	.944	.060	.157	.394	-2.732	1.381	
Z score:		Total	379	.325	.875	.045	.237	.414	-2.732	1.381	
Subdomain 4.1		Fixed Effects		.872	.045	.237	.413				
Personnel practices	Model	Random Effects			.079	.074	.576				.009
		Rural	66	.118	.772	.095	-.071	.308	-1.726	1.690	
		Small Town	26	.328	.611	.120	.082	.575	-1.043	1.463	
Z score:		Suburban	42	.086	.914	.141	-.199	.370	-2.485	1.690	
Subdomain		Urban	243	.250	.870	.056	.140	.360	-2.865	1.690	
4.2		Total	377	.214	.843	.043	.128	.299	-2.865	1.690	
Structuring and maintaining time		Fixed Effects		.844	.043	.128	.299				
	Model	Random Effects			.043	.077	.352				.001
		Rural	69	.296	.890	.107	.082	.509	-1.389	1.821	
		Small Town	25	.244	.872	.174	-.116	.603	-1.343	1.660	
Z score:		Suburban	42	.391	.787	.122	.146	.637	-1.389	1.660	
Subdomain		Urban	254	.292	.974	.061	.171	.412	-2.031	1.821	
4.3 School resources are focused on student learning		Total	390	.300	.932	.047	.207	.393	-2.031	1.821	
		Fixed Effects		.935	.047	.207	.393				
	Model	Random Effects			.0473[a]	.149[a]	.451[a]				-.011
		Rural	66	.479	.903	.111	.257	.701	-1.839	1.825	
Z score:								1.05			
Subdomain 4.4		Small Town	26	.733	.784	.154	.416	0	-1.228	1.825	
Integrating		Suburban	43	.554	.041	.159	.234	.875	-1.839	1.825	
external		Urban	245	.051	.986	.063	-.073	.175	-1.839	1.825	
expertise		Total	380	.229	.993	.051	.129	.329	-1.839	1.825	
into the school instructional program		Fixed Effects		.966	.050	.131	.326				
	Model	Random Effects			.220	-.472	.930				.099
Z score:		Rural	66	.259	.256	.754	.093	.070	.441	-1.203	
Subdomain 4.5		Small Town	26	.206	.175	.727	.143	-.118	.469	-1.335	
Coordinating		Suburban	43	.215	.280	.922	.141	-.004	.563	-1.617	
and		Urban	245	.289	.282	.941	.060	.163	.400	-2.396	
supervising		Total	380	.270	.270	.893	.046	.180	.360	-2.396	
relations with families and external communities		Fixed Effects		.896	.046	.179	.360				
	Model	Random Effects			.046[a]	.123[a]	.416[a]				-.010
		Rural	70	.569	.728	.087	.395	.742	-1.813	1.855	
Z score:		Small Town	27	.689	.781	.150	.381	.998	-.765	1.855	
Subdomain		Suburban	46	.508	.762	.112	.282	.734	-1.738	1.855	
5.1 Clear,		Urban	259	.167	.991	.062	.046	.289	-2.996	1.855	
consistent, and enforced expectations for student		Total	402	.311	.931	.046	.220	.403	-2.996	1.855	
	Model	Fixed Effects		.913	.046	.222	.401				

behavior		Random Effects			.174	-.242	.865			.061
		Rural	68	.168	.946	.115	-.061	.397	-3.430	2.031
		Small Town	26	.569	.649	.127	.307	.831	-.203	2.031
		Suburban	44	.180	.928	.140	-.103	.462	-3.430	2.031
Z score:		Urban	250	.017	1.032	.065	-.111	.146	-3.430	2.031
Subdomain		Total	388	.099	.992	.050	.000	.198	-3.430	2.031
5.2 Clean and safe		Fixed Effects		.985	.050	.001	.198			
learning		Random								
environment	Model	Effects			.118	-.275	.474			.024
Z score:		Rural	66	.256	.860	.106	.045	.468	-1.686	1.601
Subdomain		Small Town	26	.657	.670	.131	.386	.927	-1.063	1.673
5.3 Student		Suburban	42	.209	.911	.141	-.075	.493	-2.011	1.782
support services		Urban	246	.168	.984	.063	.045	.292	-2.011	2.323
provide safe		Total	380	.222	.942	.048	.127	.317	-2.011	2.323
haven for students who		Fixed Effects		.938	.048	.127	.316			
traditionally		Random								
struggle	Model	Effects			.097	-.086	.529			.014

a. Warning: Between-component variance is negative. It was replaced by 0.0 in computing this random effects measure.

Appendix M

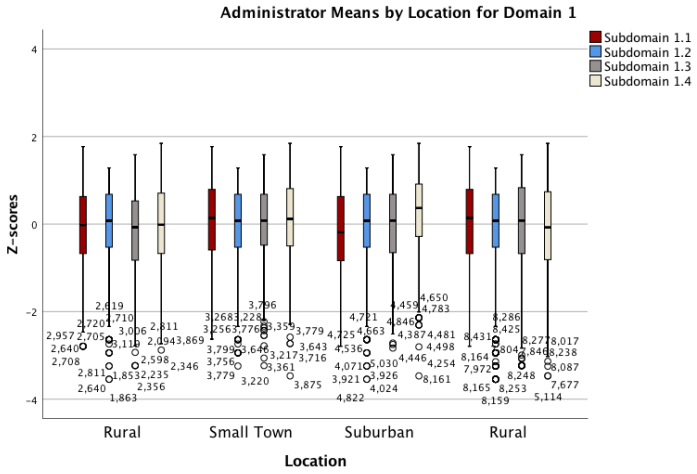
ANOVA Results

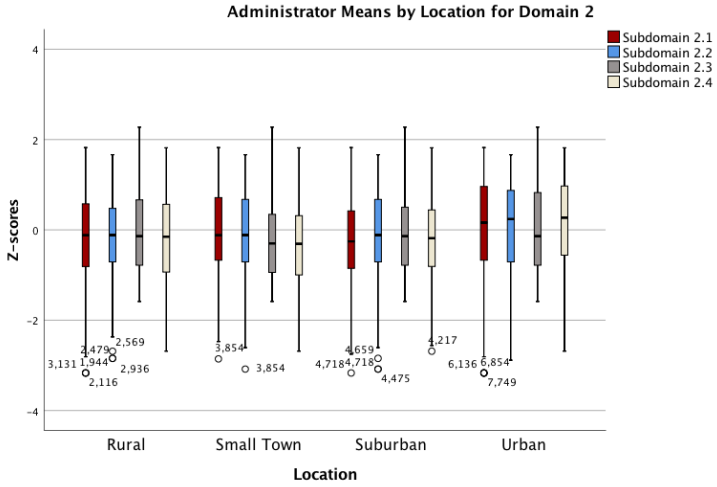
		Sum of Squares	df	Mean Square	F	p	Partial Eta Squared
Z score: Subdomain 1.1: Maintaining a school-wide focus on learning	Between Groups	2.348	3	.783	1.179	.318	.009
	Within Groups	272.273	410	.664			
	Total	274.621	413				
Z score: Subdomain 1.2 Formal leaders are recognized as instructional leaders	Between Groups	4.079	3	1.360	1.814	.144	.013
	Within Groups	302.081	403	.750			
	Total	306.161	406				
Z score: Subdomain 1.3: Collaborative design of integrated learning plan	Between Groups	2.822	3	.941	1.129	.337	.008
	Within Groups	336.609	404	.833			
	Total	339.432	407				
Z score: Subdomain 1.4 Providing appropriate services for students who traditionally struggle	Between Groups	15.690	3	5.230	5.238	.001	.038
	Within Groups	395.403	396	.998			
	Total	411.093	399				
Z score: Subdomain 2.1 Formative evaluation of student learning	Between Groups	7.694	3	2.565	2.682	.047	.020
	Within Groups	370.009	387	.956			
	Total	377.702	390				
Z score: Subdomain 2.2: Summative evaluation of student learning	Between Groups	16.032	3	5.344	6.693	.000	.048
	Within Groups	317.789	398	.798			
	Total	333.821	401				
Z score: Subdomain 2.3 Formative evaluation of teaching	Between Groups	2.155	3	.718	.809	.490	.006
	Within Groups	340.155	383	.888			
	Total	342.309	386				
Z score: Subdomain 2.4 Summative evaluation of teaching	Between Groups	5.609	3	1.870	2.655	.048	.020
	Within Groups	281.689	400	.704			
	Total	287.298	403				
Z score: Subdomain 3.1 Collaborative school-wide focus on problems of teaching and learning	Between Groups	1.317	3	.439	.480	.697	.004
	Within Groups	364.352	398	.915			
	Total	365.669	401				
Z score: Subdomain 3.2 Professional learning	Between Groups	1.992	3	.664	.778	.507	.006
	Within Groups	339.789	398	.854			
	Total	341.781	401				
Z score: Subdomain 3.3 Socially distributed leadership	Between Groups	2.844	3	.948	1.190	.313	.009
	Within Groups	317.974	399	.797			
	Total	320.818	402				
Z score: Subdomain 3.4 Coaching and Mentoring	Between Groups	1.346	3	.449	.464	.708	.004
	Within Groups	363.825	376	.968			
	Total	365.172	379				
Z score: Subdomain 4.1 Personnel practices	Between Groups	4.114	3	1.371	1.804	.146	.014
	Within Groups	285.029	375	.760			
	Total	289.143	378				
Z score: Subdomain 4.2 Structuring and maintaining time	Between Groups	1.944	3	.648	.910	.436	.007
	Within Groups	265.457	373	.712			
	Total	267.401	376				
Z score: Subdomain 4.3 School resources are	Between Groups	.450	3	.150	.171	.916	.001
	Within Groups	337.437	386	.874			

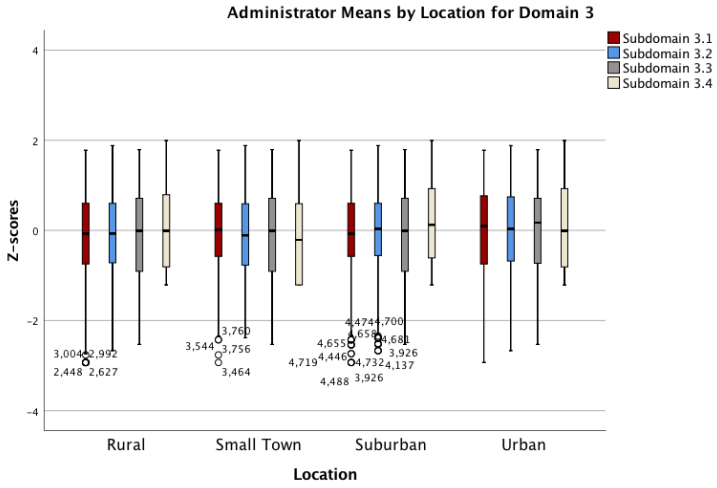
focused on student learning	Total	337.887	389				
Z score: Subdomain 4.4 Integrating external expertise into the school instructional program	Between Groups	23.085	3	7.695	8.246	.000	.062
	Within Groups	350.866	376	.933			
	Total	373.952	379				
Z score: Subdomain 4.5 Coordinating and supervising relations with families and external communities	Between Groups	.284	3	.095	.118	.950	.001
	Within Groups	301.698	376	.802			
	Total	301.981	379				
Z score: Subdomain 5.1 Clear, consistent, and enforced expectations for student behavior	Between Groups	15.652	3	5.217	6.260	.000	.045
	Within Groups	331.695	398	.833			
	Total	347.347	401				
Z score: Subdomain 5.2 Clean and safe learning environment	Between Groups	8.016	3	2.672	2.752	.042	.021
	Within Groups	372.876	384	.971			
	Total	380.892	387				
Z score: Subdomain 5.3 Student support services provide safe haven for students who traditionally struggle	Between Groups	5.703	3	1.901	2.162	.092	.017
	Within Groups	330.578	376	.879			
	Total	336.281	379				

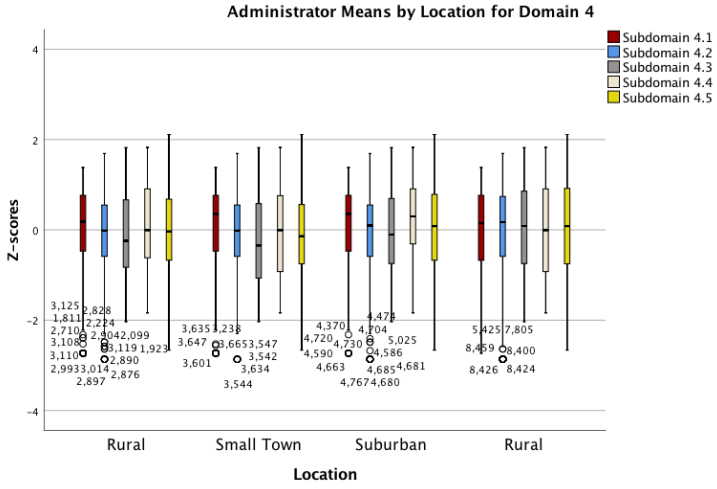
Appendix N

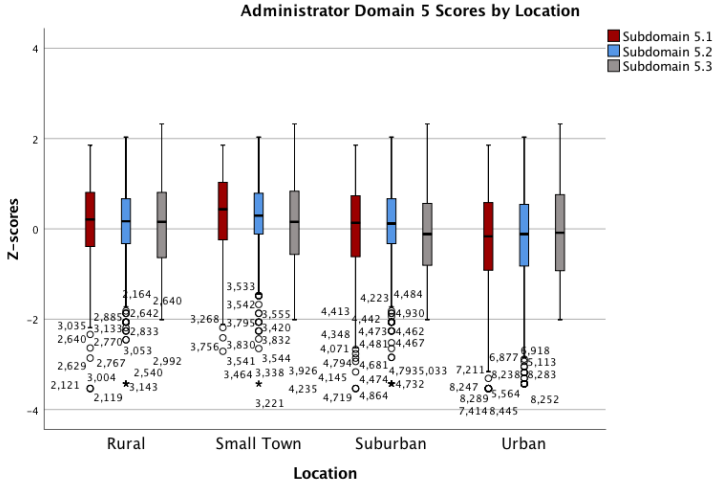
Box and Whisker Plots of the Difference in Means for Administrators by Regional Context






Administrator Means by Location for Domain 3
Subdomain 3.1
Subdomain 3.2
Subdomain 3.3
Subdomain 3.4
Z-scores
Rural
Small Town
Suburban
Urban
Location


Administrator Means by Location for Domain 4
Subdomain 4.1
Subdomain 4.2
Subdomain 4.3
Subdomain 4.4
Subdomain 4.5
Z-scores
Rural
Small Town
Suburban
Rural
Location


Administrator Domain 5 Scores by Location
Subdomain 5.1
Subdomain 5.2
Subdomain 5.3
Z-scores
Rural
Small Town
Suburban
Urban
Location

Appendix O

Post Hoc Analysis: Multiple Comparisons: Bonferroni

Dependent Variable	(I) RegionNumST	(J) RegionNumST	Mean Difference (I-J)	SE	p	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Z score: Subdomain 1.1: Maintaining a school wide focus on learning	Rural	Small Town	-.337	.181	.380	-.817	.143
		Suburban	-.107	.153	1.000	-.513	.300
		Urban	-.118	.108	1.000	-.404	.167
	Small Town	Rural	.337	.181	.380	-.143	.817
		Suburban	.231	.195	1.000	-.287	.748
		Urban	.219	.162	1.000	-.210	.648
	Suburban	Rural	.107	.153	1.000	-.300	.513
		Small Town	-.231	.195	1.000	-.748	.287
		Urban	-.012	.130	1.000	-.357	.333
	Urban	Rural	.118	.108	1.000	-.167	.404
		Small Town	-.219	.162	1.000	-.648	.210
		Suburban	.012	.130	1.000	-.333	.357
Z score: Subdomain 1.2 Formal leaders are recognized as instructional leaders	Rural	Small Town	.081	.193	1.000	-.430	.593
		Suburban	.141	.163	1.000	-.292	.575
		Urban	.254	.115	.170	-.052	.559
	Small Town	Rural	-.081	.193	1.000	-.593	.430
		Suburban	.060	.208	1.000	-.490	.610
		Urban	.172	.172	1.000	-.284	.629
	Suburban	Rural	-.141	.163	1.000	-.575	.292
		Small Town	-.060	.208	1.000	-.610	.490
		Urban	.112	.138	1.000	-.255	.479
	Urban	Rural	-.254	.115	.170	-.559	.052
		Small Town	-.172	.172	1.000	-.629	.284
		Suburban	-.112	.138	1.000	-.479	.255
Z score: Subdomain 1.3: Collaborative design of integrated learning plan	Rural	Small Town	-.373	.203	.403	-.912	.166
		Suburban	-.124	.173	1.000	-.584	.336
		Urban	-.104	.121	1.000	-.426	.218
	Small Town	Rural	.373	.203	.403	-.166	.912
		Suburban	.249	.220	1.000	-.333	.832
		Urban	.269	.181	.830	-.212	.751
	Suburban	Rural	.124	.173	1.000	-.336	.584
		Small Town	-.249	.220	1.000	-.832	.333
		Urban	.020	.147	1.000	-.370	.410
	Urban	Rural	.104	.121	1.000	-.218	.426
		Small Town	-.269	.181	.830	-.751	.212
		Suburban	-.020	.147	1.000	-.410	.370
Z score: Subdomain 1.4 Providing appropriate services for students who traditionally struggle	Rural	Small Town	-.246	.227	1.000	-.848	.355
		Suburban	-.405	.191	.211	-.912	.103
		Urban	.172	.135	1.000	-.187	.531
	Small Town	Rural	.246	.227	1.000	-.355	.848
		Suburban	-.159	.243	1.000	-.804	.486
		Urban	.418	.202	.236	-.118	.954
	Suburban	Rural	.405	.191	.211	-.103	.912
		Small Town	.159	.243	1.000	-.486	.804
		Urban	.577*	.161	.002	.149	1.005

	Urban	Rural	-.172	.135	1.000	-.531	.187
		Small Town	-.418	.202	.236	-.954	.118
		Suburban	-.577*	.161	.002	-.005	-.149
Z score: Subdomain 2.1 Formative evaluation of student learning	Rural	Small Town	-.197	.222	1.000	-.787	.393
		Suburban	.212	.189	1.000	-.289	.714
		Urban	-.202	.134	.790	-.556	.152
	Small Town	Rural	.197	.222	1.000	-.393	.787
		Suburban	.409	.239	.525	-.224	1.043
		Urban	-.005	.198	1.000	-.530	.520
	Suburban	Rural	-.212	.189	1.000	-.714	.289
		Small Town	-.409	.239	.525	-1.043	.224
		Urban	-.414	.160	.059	-.838	.009
	Urban	Rural	.202	.134	.790	-.152	.556
		Small Town	.005	.198	1.000	-.520	.530
		Suburban	.414	.160	.059	-.009	.838
Z score: Subdomain 2.2: Summative evaluation of student learning	Rural	Small Town	-.139	.202	1.000	-.676	.398
		Suburban	.162	.170	1.000	-.288	.612
		Urban	-.369*	.120	.014	-.689	-.050
	Small Town	Rural	.139	.202	1.000	-.398	.676
		Suburban	.301	.217	.992	-.273	.875
		Urban	-.230	.181	1.000	-.710	.249
	Suburban	Rural	-.162	.170	1.000	-.612	.288
		Small Town	-.301	.217	.992	-.875	.273
		Urban	-.531*	.143	.001	-.911	-.152
	Urban	Rural	.369*	.120	.014	.050	.689
		Small Town	.230	.181	1.000	-.249	.710
		Suburban	.531*	.143	.001	.152	.911
Z score: Subdomain 2.3 Formative evaluation of teaching	Rural	Small Town	-.195	.217	1.000	-.771	.381
		Suburban	.153	.185	1.000	-.338	.643
		Urban	.057	.129	1.000	-.284	.399
	Small Town	Rural	.195	.217	1.000	-.381	.771
		Suburban	.348	.235	.840	-.276	.971
		Urban	.252	.194	1.000	-.262	.767
	Suburban	Rural	-.153	.185	1.000	-.643	.338
		Small Town	-.348	.235	.840	-.971	.276
		Urban	-.095	.157	1.000	-.512	.321
	Urban	Rural	-.057	.129	1.000	-.399	.284
		Small Town	-.252	.194	1.000	-.767	.262
		Suburban	.095	.157	1.000	-.321	.512
Z score: Subdomain 2.4 Summative evaluation of teaching	Rural	Small Town	-.070	.187	1.000	-.566	.427
		Suburban	.223	.160	.979	-.201	.647
		Urban	-.143	.112	1.000	-.441	.155
	Small Town	Rural	.070	.187	1.000	-.427	.566
		Suburban	.293	.202	.885	-.242	.829
		Urban	-.073	.167	1.000	-.516	.369
	Suburban	Rural	-.223	.160	.979	-.647	.201
		Small Town	-.293	.202	.885	-.829	.242
		Urban	-.366*	.135	.043	-.726	-.007
	Urban	Rural	.143	.112	1.000	-.155	.441
		Small Town	.073	.167	1.000	-.369	.516
		Suburban	.366*	.135	.043	.007	.726
Z score: Subdomain	Rural	Small Town	-.250	.217	1.000	-.825	.325
		Suburban	-.031	.183	1.000	-.516	.454

3.1 Collaborative school-wide focus on problems of teaching and learning	Small Town	Urban	-.082	.129	1.000	-.423	.260
		Rural	.250	.217	1.000	-.325	.825
		Suburban	.219	.233	1.000	-.399	.837
	Suburban	Urban	.168	.193	1.000	-.345	.681
		Rural	.031	.183	1.000	-.454	.516
		Small Town	-.219	.233	1.000	-.837	.399
	Urban	Urban	-.051	.154	1.000	-.460	.359
		Rural	.082	.129	1.000	-.260	.423
		Small Town	-.168	.193	1.000	-.681	.345
		Suburban	.051	.154	1.000	-.359	.460
Z score: Subdomain 3.2 Professional learning	Rural	Small Town	-.251	.209	1.000	-.806	.304
		Suburban	.056	.177	1.000	-.412	.525
		Urban	-.085	.124	1.000	-.415	.245
	Small Town	Rural	.251	.209	1.000	-.304	.806
		Suburban	.307	.225	1.000	-.289	.903
		Urban	.166	.187	1.000	-.330	.661
	Suburban	Rural	-.056	.177	1.000	-.525	.412
		Small Town	-.307	.225	1.000	-.903	.289
		Urban	-.141	.149	1.000	-.537	.254
	Urban	Rural	.085	.124	1.000	-.245	.415
		Small Town	-.166	.187	1.000	-.661	.330
		Suburban	.141	.149	1.000	-.254	.537
Z score: Subdomain 3.3 Socially distributed leadership	Rural	Small Town	-.354	.202	.485	-.890	.182
		Suburban	.013	.169	1.000	-.436	.462
		Urban	-.075	.120	1.000	-.394	.244
	Small Town	Rural	.354	.202	.485	-.182	.890
		Suburban	.367	.216	.545	-.207	.941
		Urban	.279	.181	.738	-.200	.758
	Suburban	Rural	-.013	.169	1.000	-.462	.436
		Small Town	-.367	.216	.545	-.941	.207
		Urban	-.088	.143	1.000	-.467	.291
	Urban	Rural	.075	.120	1.000	-.244	.394
		Small Town	-.279	.181	.738	-.758	.200
		Suburban	.088	.143	1.000	-.291	.467
Z score: Subdomain 3.4 Coaching and Mentoring	Rural	Small Town	-.162	.228	1.000	-.766	.442
		Suburban	.015	.194	1.000	-.500	.530
		Urban	-.120	.136	1.000	-.481	.242
	Small Town	Rural	.162	.228	1.000	-.442	.766
		Suburban	.177	.245	1.000	-.474	.828
		Urban	.043	.203	1.000	-.495	.581
	Suburban	Rural	-.015	.194	1.000	-.530	.500
		Small Town	-.177	.245	1.000	-.828	.474
		Urban	-.135	.164	1.000	-.570	.301
	Urban	Rural	.120	.136	1.000	-.242	.481
		Small Town	-.043	.203	1.000	-.581	.495
		Suburban	.135	.164	1.000	-.301	.570
Z score: Subdomain 4.1 Personnel practices	Rural	Small Town	-.299	.202	.837	-.834	.236
		Suburban	-.243	.172	.951	-.700	.213
		Urban	.007	.121	1.000	-.314	.327
	Small Town	Rural	.299	.202	.837	-.236	.834
		Suburban	.056	.218	1.000	-.521	.633
		Urban	.306	.180	.540	-.171	.783
	Suburban	Rural	.243	.172	.951	-.213	.700
		Small Town	-.056	.218	1.000	-.633	.521

		Urban	.250	.146	.522	-.136	.636
	Urban	Rural	-.007	.121	1.000	-.327	.314
		Small Town	-.306	.180	.540	-.783	.171
		Suburban	-.250	.146	.522	-.636	.136
Z score: Subdomain 4.2 Structuring and maintaining time	Rural	Small Town	-.210	.195	1.000	-.728	.308
		Suburban	.033	.167	1.000	-.409	.474
		Urban	-.131	.117	1.000	-.442	.179
	Small Town	Rural	.210	.195	1.000	-.308	.728
		Suburban	.243	.211	1.000	-.316	.801
		Urban	.078	.174	1.000	-.383	.540
	Suburban	Rural	-.033	.167	1.000	-.474	.409
		Small Town	-.243	.211	1.000	-.801	.316
		Urban	-.164	.141	1.000	-.538	.210
	Urban	Rural	.131	.117	1.000	-.179	.442
		Small Town	-.078	.174	1.000	-.540	.383
		Suburban	.164	.141	1.000	-.210	.538
Z score: Subdomain 4.3 School resources are focused on student learning	Rural	Small Town	.052	.218	1.000	-.527	.631
		Suburban	-.096	.183	1.000	-.581	.389
		Urban	.004	.127	1.000	-.333	.340
	Small Town	Rural	-.052	.218	1.000	-.631	.527
		Suburban	-.148	.236	1.000	-.774	.478
		Urban	-.048	.196	1.000	-.568	.472
	Suburban	Rural	.096	.183	1.000	-.389	.581
		Small Town	.148	.236	1.000	-.478	.774
		Urban	.100	.156	1.000	-.313	.513
	Urban	Rural	-.004	.127	1.000	-.340	.333
		Small Town	.048	.196	1.000	-.472	.568
		Suburban	-.100	.156	1.000	-.513	.313
Z score: Subdomain 4.4 Integrating external expertise into the school instructional program	Rural	Small Town	-.254	.224	1.000	-.847	.339
		Suburban	-.075	.189	1.000	-.577	.427
		Urban	.428*	.134	.009	.073	.784
	Small Town	Rural	.254	.224	1.000	-.339	.847
		Suburban	.179	.240	1.000	-.458	.815
		Urban	.683*	.199	.004	.154	1.211
	Suburban	Rural	.075	.189	1.000	-.427	.577
		Small Town	-.179	.240	1.000	-.815	.458
		Urban	.504*	.160	.010	.080	.927
	Urban	Rural	-.428*	.134	.009	-.784	-.073
		Small Town	-.683*	.199	.004	-.211	-.154
		Suburban	-.504*	.160	.010	-.927	-.080
Z score: Subdomain 4.5 Coordinating and supervising relations with families and external communities	Rural	Small Town	.080	.207	1.000	-.470	.630
		Suburban	-.024	.176	1.000	-.489	.442
		Urban	-.026	.124	1.000	-.355	.304
	Small Town	Rural	-.080	.207	1.000	-.630	.470
		Suburban	-.104	.223	1.000	-.694	.486
		Urban	-.106	.185	1.000	-.596	.384
	Suburban	Rural	.024	.176	1.000	-.442	.489
		Small Town	.104	.223	1.000	-.486	.694
		Urban	-.002	.148	1.000	-.395	.391
	Urban	Rural	.026	.124	1.000	-.304	.355
		Small Town	.106	.185	1.000	-.384	.596
		Suburban	.002	.148	1.000	-.391	.395
Z score: Subdomain	Rural	Small Town	-.121	.207	1.000	-.669	.428
		Suburban	.061	.173	1.000	-.399	.520

5.1 Clear, consistent, and enforced expectations for student behavior	Small Town	Urban	.401*	.123	.007	.075	.728
		Rural	.121	.207	1.000	-.428	.669
		Suburban	.181	.221	1.000	-.406	.768
	Suburban	Urban	.522*	.185	.030	.032	1.012
		Rural	-.061	.173	1.000	-.520	.399
		Small Town	-.181	.221	1.000	-.768	.406
	Urban	Urban	.341	.146	.121	-.046	.728
		Rural	-.401*	.123	.007	-.728	-.075
		Small Town	-.522*	.185	.030	-1.012	-.032
		Suburban	-.341	.146	.121	-.728	.046
Z score: Subdomain 5.2 Clean and safe learning environment	Rural	Small Town	-.401	.227	.471	-1.003	.202
		Suburban	-.012	.191	1.000	-.517	.494
		Urban	.151	.135	1.000	-.207	.508
	Small Town	Rural	.401	.227	.471	-.202	1.003
		Suburban	.389	.244	.666	-.257	1.036
		Urban	.551*	.203	.041	.013	1.090
	Suburban	Rural	.012	.191	1.000	-.494	.517
		Small Town	-.389	.244	.666	-1.036	.257
		Urban	.162	.161	1.000	-.265	.589
	Urban	Rural	-.151	.135	1.000	-.508	.207
		Small Town	-.551*	.203	.041	-.090	-.013
		Suburban	-.162	.161	1.000	-.589	.265
Z score: Subdomain 5.3 Student support services provide safe haven for students who traditionally struggle	Rural	Small Town	-.400	.217	.396	-.976	.176
		Suburban	.047	.185	1.000	-.444	.538
		Urban	.088	.130	1.000	-.257	.433
	Small Town	Rural	.400	.217	.396	-.176	.976
		Suburban	.448	.234	.339	-.173	1.068
		Urban	.488	.193	.072	-.025	1.001
	Suburban	Rural	-.047	.185	1.000	-.538	.444
		Small Town	-.448	.234	.339	-1.068	.173
		Urban	.041	.157	1.000	-.375	.456
	Urban	Rural	-.088	.130	1.000	-.433	.257
		Small Town	-.488	.193	.072	-1.001	.025
		Suburban	-.041	.157	1.000	-.456	.375

*. The mean difference is significant at the 0.05 level.

Appendix P

Copy of Institutional Review Board Exempt Status

TEACHERS COLLEGE
COLUMBIA UNIVERSITY

Teachers College IRB *Exempt Study Approval*

To: Margaret Freeman
From: Myra Luna Lucero, Research Compliance Manager
Subject: IRB Approval: 20-138 Protocol
Date: 12/30/2019

Thank you for submitting your study entitled, "*The Difference in Distributed Principal Leadership Practices Across Context According to the CALL*;" the IRB has determined that your study is **Exempt** from committee review (Category **4**) on 12/30/2019.

Please keep in mind that the IRB Committee must be contacted if there are any changes to your research protocol. The number assigned to your protocol is **20-138**. Feel free to contact the IRB Office by using the "Messages" option in the electronic Mentor IRB system if you have any questions about this protocol.

As the PI of record for this protocol, you are required to:

- Use current, up-to-date IRB approved documents
- Ensure all study staff and their CITI certifications are on record with the IRB
- Notify the IRB of any changes or modifications to your study procedures
- Alert the IRB of any adverse events

You are also required to respond if the IRB communicates with you directly about any aspect of your protocol. Failure to adhere to your responsibilities as a study PI can result in action by the IRB up to and including suspension of your approval and cessation of your research.

You can retrieve a PDF copy of this approval letter from Mentor IRB.

Best wishes for your research work.

Sincerely,
Dr. Myra Luna Lucero
Research Compliance Manager
IRB@tc.edu

Appendix Q

Research Brief for the CALL Team: Furthering CALL Research

FURTHERING CALL RESEARCH

MEG FREEMAN

TEACHERS COLLEGE, COLUMBIA UNIVERSITY

Leadership for learning has arisen as a robust framework for measuring principal leadership practices that impact student achievement by synthesizing fundamental theories of principal leadership, such as instructional leadership, transformational leadership, and shared instructional leadership. The purpose of this study is to examine how administrators and teachers prioritize leadership for learning practices differently. These differences will allow us to identify opportunities for preparation, support, and professional development for principals. Collected by the researchers of the Comprehensive Assessment of Leadership for Learning (CALL) group at the University of Wisconsin-Madison, I use the 2016-2019 CALL dataset in the present study, which includes 281 schools with 536 administrators and 8,027 teachers.

This research extends the work of Blitz and Modeste (2015) by using the same methods to compare differences in principal leadership between teachers and administrators. This is important for two reasons. First, the results affirm Blitz and Modeste's finding that teacher and administrator responses on the CALL survey are different. Since a common definition of leadership is influence over followers, this finding shows that current evaluation systems may be missing a key piece of data that could be used for improvement: how teachers view their administrator's leadership (Boyce & Bowers, 2018b; Reich; 1988; Urick & Bowers, 2019). Second, comparing this current CALL data set to the Blitz and Modeste (2015) data confirms previous data the principal leadership is complex and varied depending on the context and systems designed for principal evaluation and growth should treat it as such (Halverson & Kelly, 2017; Leithwood & Jantzi, 2008; Marks & Printy, 2003; Printy et al., 2009).

WHY the CALL? The CALL is used for this study as it is built on research that highlights leadership tasks that have been proved through research to have an impact on student achievement. For the CALL survey, these leadership tasks have been corralled under five domains: focus on learning, monitoring teaching and learning, building a professional community, acquiring and allocating resources, and establishing a safe and effective learning environment. These domains are expanded through twenty-subdomains

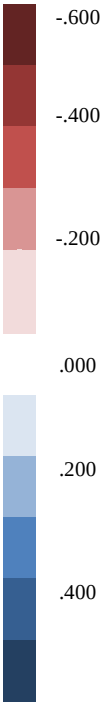
Following the previous research using the CALL survey data, the survey items were averaged by subdomains to create a subdomain score for each survey participant to create the twenty subdomains (Blitz & Modeste, 2015; Bowers et al., 2017). These twenty subdomain scores were converted to z-scores to standardized responses for both groups of respondent. Twenty independent samples t-test were run at $p = <.05$ with a FWER of $p = <.025$ on each of the twenty subdomains of CALL to determine the difference between the means of teachers and the means of principals for each domain. Effect sizes were calculated for each mean difference using Cohen's D.

To compare the Blitz and Modeste data with the 2016-2019 CALL dataset, a table with heat gradient is used in the tables to the right to show the effect size differences between teacher and administrator means. There is a consistent structure in the table where darker red or blue indicates a larger difference in the responses. A red square illustrates a negative effect size or mean difference, showing that the administrators have the higher score. A blue square shows a positive effect size or mean difference illustrating that teachers have the higher score. Tables should be read left to right.

Table 1

Table of Cohen's d Effect Sizes of Differences Between Teachers and Administrators by Subdomain for 2016-2019 CALL Dataset

2016-2019 CALL DATASET DOMAINS AND SUBDOMAINS				
1: Focus on Learning	*2: Monitoring Teaching & Learning*	*3:Building Nested Learning Communities*	*4: Acquiring & Allocating Resources*	*5: Maintaining Learning Environment*
1.1 School-wide focus - learning (ES = -.437)*	2.1 Formative evaluation - students (ES = -.201)*	3.1 School-wide focus - problems (ES = -.276)*	4.1 Personnel practices (ES = -.332)*	5.1 Expectations for student behavior (ES = -.339)*
1.2 Instructional leaders (ES = -.393)*	2.2 Summative evaluation - students (ES = -.185)*	3.2 Professional learning (ES = -.273)*	4.2 Structuring and maintaining time (ES = -.212)*	5.2 Safe learning environment (ES = -.154)*
1.3 Integrated learning plan (ES = -.148)*	2.3 Formative evaluation - teaching (ES = -.373)*	3.3 Socially distributed leadership (ES = -.241)*	4.3 School resources (ES = -.354)*	5.3 Student support services (ES = -.243)*
1.4 Providing appropriate services (ES = -.084)	2.4 Summative evaluation - teaching (ES = -.368)*	3.4 Coaching and mentoring (ES = -.183)*	4.4 Integrating external expertise (ES = -.222)*	
			4.5 Relations with families & external communities (ES = -.324)*	



Note: *Significant p values at p = <.05 with a FWER of p = <.0025. This table is presented in color. A larger negative effect size (deeper red) indicates that administrator means are larger. A larger positive effect size (deeper blue) indicates that teacher means are larger

Table 1 contains the results of the comparison between teacher and administrator scores in the 2016-2019 CALL dataset. The second table, Table 2 contains the Blitz and Modeste (2015) data. The Blitz and Modeste (2015) data are not new data; As illustrated in the tables of Tables 1 and 2, a comparison of the 2016-2019 CALL dataset to the Blitz and Modeste (2015) data shows that the biggest differences in subdomains between teachers and administrators do not replicate between datasets. The differences between datasets in subdomain scores could be for several different reasons. First, it could be random error as the samples are different. Second, it could be because the 2016-2019 CALL dataset is larger, as the Blitz and Modeste (2015) data is half the size of the 2016-2019 CALL datset. Third, the samples could be different. However, there is not enough data on the demographics of the Blitz and Modeste (2015) study to determine if the context of the two datasets is, in fact, significantly different. The significant finding in this comparison is that teachers and administrators have statistically different means in nearly all of the subdomains on both data sets, which has implications for how leaders are assessed and supported as in two samples, teachers and administrators rated leadership tasks differently. Similarly, looking at particular subdomains that have similar large differences between both groups provide an opportunity to examine the difficulty of those specific leadership tasks and why there is a difference in scores.

In both the largest and the smallest differences, there is a consistent subdomain across the 2016-2019 CALL dataset and the Blitz and Modeste (2015) data. The subdomain, *Formal leaders are recognized as instructional leaders* (1.2), shows a consistent large difference between both datasets. This area is unique to the survey in that it is the only question that focuses on the actions of a school's principal. Similarly, the subdomain, *providing appropriate services for students who traditionally struggle* (1.4), shows a small difference between both datasets. This subdomain is unique in that it focuses on the teacher's leadership actions within the classroom. The purpose of the formative nature of the CALL is for survey participants to self-reflect and improve their practice. These areas seem to illustrate rating inflation where both administrators and teachers rated themselves high in areas where they had primary responsibility.

Table 2

Table of Mean Differences Between Teachers and Administrators by Subdomain for Blitz and Modeste (2015) Data

BLITZ & MODESTE (2015) DOMAINS AND SUBDOMAINS				
1: Focus on Learning	*2: Monitoring Teaching & Learning*	*3:Building Nested Learning Communities*	*4: Acquiring & Allocating Resources*	*5: Maintaining Learning Environment*
1.1 School-wide focus - learning (MD = -.379)*	2.1 Formative eval-students (MD = -.056)*	3.1 School-wide focus - problems (MD = -.437)*	4.1 Personnel practices (MD = -.302)*	5.1 Expectations for student behavior (MD =-.386)*
1.2 Instructional leaders (MD = -.674)*	2.2 Summative eval-students (MD = -.053)*	3.2 Professional learning (MD = -.215)*	4.2 Structuring and maintaining time (MD=-.239)*	5.2 Safe learning environment (MD = -.502)**
1.3 Integrated learning plan (MD = -.393)*	2.3 Formative eval–teaching (MD = -.434)	3.3 Socially distributed leadership (MD = -.591)*	4.3 School resources (MD = -.419)*	5.3 Student support services (MD = -.217)*
1.4 Providing appropriate services (MD = .163)*	2.4 Summative eval-teaching (MD = -.259)	3.4 Coaching and mentoring (MD = -.225)*	4.4 Integrating external expertise (MD = -.412)*	5.4 Buffering the teaching environment (MD = -.400)*
			4.5 Relations with families & external communities (MD = -.274)*	

-.600 -.400 -.200 .000 .200 .400

Note: Data and analysis using the CALL 2012 sample adapted from Blitz and Modeste (2015) *Significant p – values. This table is presented in color. A larger negative effect size (deeper red) indicates that administrator means are larger. A larger positive effect size (deeper blue) indicates that teacher means are larger

In comparing the 2016-2019 dataset to the Blitz and Modeste (2015) data, this analysis indicates that the biggest differences in subdomains between teachers and administrators do not replicate between datasets. The difference could be due to changing perceptions over time, or it could simply be that they are different data sets. Because of this, the comparison of the two groups does not have substantial implications for the understanding of how context affects leadership for learning practices. However, the difference between the groups may indicate that any two different samples of leaders would prioritize tasks differently. The difference might show the need for principal evaluations to be more aligned to district strategic goals as opposed to generic state rubrics (Close et al., 2018; Corcoran et al., 2013; Leithwood et al., 2004; Sun & Youngs, 2009). The principals' perceptions of leadership and leadership strengths varied greatly across datasets, indicating that levers for principal growth, such as evaluation and professional development, ought to be more specific to the differentiated needs of those school leaders (Bowers et al., 2017). These differences also point to the need to move to a more developmental system of professional growth where principals build on these goals over time.

Implications

- The difference in scores between the Blitz and Modeste (2015) sample and the current sample affirms the need for principal evaluations to be aligned to district strategic goals and focus on instructional leadership activities as opposed to a generic set of tasks in order to capture this adaptive nature of leadership (Close et al., 2018; Corcoran et al., 2013; Leithwood et al., 2004; Sun & Youngs, 2009). Principals' perceptions of leadership and leadership strengths varied greatly, indicating that levers for principal growth, such as evaluation and professional development, ought to be more varied to match the differentiated needs of those school leaders (Bowers et al., 2017).

- The large amount of domains with statistical difference affirms the need for evaluation systems that include a 360 degree framework. A 360 degree framework is an evaluation system that in which administrators receive feedback not only from their supervisor, but also from teachers (Goldring et al., 2015; Liu et al., 2014; Shoaf et al., 2013; Tran & Bon, 2015). Administrators are evaluated on the same characteristics by each group and then those results are compared. Leaders rate themselves very differently than teachers rate them. Leaders may believe these important aspects of leadership are being clearly enacted in a school while teachers may perceive something different. In order to improve leadership practice, feedback from teachers on how they view those leadership practices being enacted is critical.

- Another practical implication from this study is for state departments to use the data from the CALL to create focus groups or professional learning communities between different groups of principals. The differences in the two samples above show that principals rate themselves higher on different domains. If groups of principals were able to compare their results with other principals, they would be able to share strengths and weaknesses in a critical friends network with the purpose of growth. This would provide on-going, job-embedded professional growth and development for principals.

References:

Freeman, M. (2020). *The Difference in Distributed Principal Leadership Practices Across Context According to the Comprehensive Assessment of Leadership for Learning*. [Doctoral Dissertation, Columbia University, Teachers College]. ProQuest Dissertations Publishing.

Halverson, R., & Kelley, C. (2017). *Mapping leadership: The tasks that matter for improving teaching and learning in schools*: John Wiley & Sons.

Suggested Citation:

Freeman, M. (2020). *Furthering CALL Research*. CALL Research Group at Wisconsin Center for Education Products and Services.

Contact info: maf2284@tc.columbia.edu

Appendix R

Research Brief for the CALL Team: The Differences in Principal Leadership Practices Due to Context

THE DIFFERENCES IN PRINCIPAL LEADERSHIP PRACTICES DUE TO CONTEXT

MEG FREEMAN
TEACHERS COLLEGE, COLUMBIA UNIVERSITY

Leadership for learning has arisen as a robust framework for measuring principal leadership practices that impact student achievement by synthesizing fundamental theories of principal leadership, such as instructional leadership, transformational leadership, and shared instructional leadership. The purpose of this study is to examine how administrators and teachers prioritize leadership for learning practices differently. These differences will allow us to identify opportunities for preparation, support, and professional development for principals. Collected by the researchers of the Comprehensive Assessment of Leadership for Learning (CALL) group at the University of Wisconsin-Madison, I use the 2016-2019 CALL dataset in the present study, which includes 281 schools with 536 administrators and 8,027 teachers.

This study examines context in two ways: grade level and region. Because size of school and the level of pedagogy were found to be important aspects that mitigate the effect size of leadership, it is important to measure context as related to elementary or secondary schools (Hallinger, 2005; Leithwood & Jantzi, 2008; Nguni et al., 2006; Thoonen et al., 2011). Research indicates a difference in principal leadership and principal evaluation systems due especially to the context of urban schools (Fuller et al., 2015; Greenlee & Brown, 2009; Horsford et al., 2018; Hughes et al., 2015; Hunter & Donahoo, 2003; Ingersoll & May, 2016; Johnson et al., 2012; White et al., 2012). The context of region will be examined through disaggregating the data by rural, small town, suburban, or urban. Districts will be separated into these four groups to examine how their school leader scores differ from each other.

WHY the CALL? The CALL was chosen for this study as it is built on research that highlights leadership tasks that have been proved through research to have an impact on student achievement. For the CALL survey, these leadership tasks have been corralled under five domains: focus on learning, monitoring teaching and learning, building a professional community, acquiring and allocating resources, and establishing a safe and effective learning environment. These domains are expanded through twenty-subdomains.

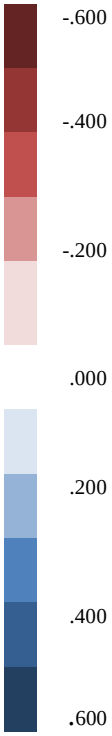
Following the previous research using the CALL survey data, the survey items were averaged by subdomains to create a subdomain score for each survey participant to create the twenty subdomains (Blitz & Modeste, 2015; Bowers et al., 2017). These twenty subdomain scores were converted to z-scores to standardized responses for both groups of respondent. To determine the difference between elementary and secondary principals, twenty independent samples t-test were run at $p = <.05$ with a FWER of $p = <.025$ on each of the twenty subdomains of CALL. Effect sizes were calculated for each mean difference using Cohen's D. To determine the difference in means due to location, a one-way analysis of variance (ANOVA) was run to investigate differences in mean scores on CALL domains by location between four categories: rural, small town, suburban, urban. Bonferroni's Post Hoc test was conducted to determine between which groups the difference occurred for each subdomain.

ADMINISTRATOR SCORES BY GRADE LEVEL

Table 1

Table with Heat Gradient of Cohen's d Effect Sizes by Subdomain for Elementary Administrator Means vs. Secondary Administrator Means

ELEMENTARY DOMAINS AND SUBDOMAINS				
1: Focus on Learning	2: Monitoring Teaching & Learning	3:Building Nested Learning Communities	4: Acquiring & Allocating Resources	5: Maintaining Learning Environment
1.1 School-wide focus - learning (ES= .001)	2.1 Formative evaluation - students (ES = .269)	3.1 School-wide focus - problems (ES = .062)	4.1 Personnel practices (ES = -.049)	5.1 Expectations for student behavior (ES =-.095)
1.2 Instructional leaders (ES =-.014)	2.2 Summative evaluation – students (ES = .457)*	3.2 Professional learning (ES = -.041)	4.2 Structuring & maintaining time (ES = .147)	5.2 Safe learning environment (ES = -.178)
1.3 Integrated learning plan (ES = -.057)	2.3 Formative evaluation - teaching (ES = -.006)	3.3 Socially distributed leadership (ES = -.138)	4.3 School resources (ES = -.059)	5.3 Student support services (ES = -.173)
1.4 Providing appropriate services (ES= -.203)	2.4 Summative evaluation - teaching (ES = .119)	3.4 Coaching and mentoring (ES = .062)	4.4 Integrating external expertise (ES = -.220)	
			4.5 Relations with families and external communities (ES = -.112)	



Note: *Significant p values at p = <.05 with a FWER of p = <.0025.

As demonstrated in Table 1, I find that in comparing the mean responses of elementary to secondary administrators across the 20 CALL subdomains, only one is statistically significant, *Summative evaluation of student learning* (2.2), t (385.861)= 4.933; p<0.001.

To illustrate effect sizes, a table with heat gradient is used in Table 1. This table provides a visual representation of the data. Secondary administrators are used as the reference group for this data set. Elementary administrator means were subtracted from secondary administrator means in the comparison. As a result, effect sizes that are negative will show that secondary administrators had the higher mean. Negative effect sizes

In looking at Table 1, two findings are visually notable. First, nearly half of the boxes (nine total) are white, indicating that there is an non-significant effect size between the means of elementary and secondary administrators on these domains. However, of the boxes that do contain color, some contain gradients of blue, and some contain gradients of red. The color illustrates where there are significant differences in mean scores between elementary and secondary administrators, in six subdomains secondary administrator means are higher and in four subdomains elementary means are higher. The one statistically significant domain, *Summative evaluation of student learning* (2.2) (ES = .457, p = <.001), had a moderate effect size, as illustrated in Table 1 through the dark blue square, meaning that elementary administrators have higher means. Between the four subdomains where there is any type of effect in this analysis, some subdomain means are higher for elementary administrators and others for secondary. In two subdomains, *providing appropriate services for students who traditionally struggle* (1.4) and *integrating external expertise into school instructional program* (4.4), the secondary administrators had higher means. Still, in *formative evaluation of student learning* (2.1) and *summative evaluation of student learning* (2.2), elementary administrators had higher means. While overall, the differences between elementary and secondary administrators are not significant, these four subdomains provide opportunities for collaboration around principal growth where administrators may be able to learn from each other.

Table 2

Table with Heat Gradient with Comparison of Administrator Subdomains with Significant Mean Differences Between Urban, Suburban, Small Town, and Rural

CALL SUBDOMAINS			
1.4 Providing appropriate services for students who traditionally struggle	Suburban vs. Urban* (ES = -.575)	Small Town vs. Urban (ES = .398)	Rural vs. Urban (ES = .168)
2.2: Summative evaluation of student learning	Suburban vs. Urban* (ES = -.587)	Small Town vs. Urban (ES= -.259)	Rural vs. Urban* (ES = -.417)
2.4: Summative evaluation of teaching	Suburban vs. Urban* (ES = -.419)	Small Town vs. Urban (ES= -.084)	Rural vs. Urban (ES= -.164)
4.4: Integrating external expertise into the school instructional program	Suburban vs. Urban* (ES = .506)	Small Town vs. Urban* (ES = .704)	Rural vs. Urban* (ES = .442)
5.1: Clear, consistent, enforced expectations for student behavior	Suburban vs. Urban (ES = .355)	Small Town vs. Urban* (ES = .536)	Rural vs. Urban* (ES = .427)
5.2: Clean and safe learning environment	Suburban vs. Urban (ES =.160)	Small Town vs. Urban* (ES = .550)	Rural vs. Urban (ES =.149)

Note: This table is presented in color. Every .10 of an increase in effect size changes the gradient of the square to be one shade darker. Effect sizes are listed in parenthesis. A larger negative effect size (deeper red) indicates that urban administrator means are larger. A larger positive effect size (deeper blue) indicates that the rural, small town, or suburban means are larger. *Significant *p* values.

A one-way analysis of variance (ANOVA) was conducted to investigate differences in mean scores on CALL domains by location between the four categories: Rural, Small Town, Suburban, Urban. ANOVA results show a significant main effect for urbancity in six subdomains (listed in Table 2). Estimates of effect sizes in the other fourteen domains revealed low strength in associations. Overall, the analysis revealed that there is about around 5% of the variance in the means for most domain averages that is due to urbanicity.

Though urban administrators have the most difference in means, this is not because they are always lower than the mean, as illustrated in Table 2. For *summative evaluation of student learning* (2.2) and *summative evaluation of teaching* (2.4), the effect size is negative, indicating that urban administrators have higher subdomain means. Conversely, in *Providing appropriate services for students who traditionally struggle* (1.4), *Summative evaluation of student learning* (2.2), *Summative evaluation of teaching* (2.4), *Integrating external expertise into the school instructional program* (4.4), *Clear, consistent, enforced expectations for student behavior* (5.1), and *Clean and safe learning environment* (5.2), the effect size is positive as the rural, small town, or suburban administrators have higher subdomain means. The results show that urban administrators score distinctly different in their perceptions of their leadership practice around leadership for learning. While there is much qualitative research on urban leadership requiring different skills, this study provides quantitative results that urban principals prioritize leadership differently around student discipline, resources, and summative evaluation (Khalifa et al., 2016; Theoharis, 2007).

Implications

- Most current evaluation models are comprised of an evaluation rubric and a data component (Close, Amrein-Beardsley, & Collins, 2018; Ehlert, Koedel, Parsons, & Podgursky, 2016; Orr, 2013; Sun & Youngs, 2009). Previous literature has shown that the data used for these methods of Value Added Models is statistically unsound, making the comparison biased, particularly against low performing schools (Close et al., 2018; Jason A. Grissom & Susanna Loeb, 2011; Orr, 2013). The data from this study suggests that the rubric component of evaluation systems may also be biased as urban principals rate themselves distinctly different from the other three groups in research-based leadership behaviors. These results show *not* that urban school principals are always underperforming, but that they are distinctly different in the way that they prioritize leadership. Changes to evaluation systems need to be made to make them more equitable across contexts of urbanicity, which might include more goals-based system that is tailored to each unique location.

- While preparation programs do not need to distinguish by grade level based on this research, a distinction between urban schools and the other locations is needed. Principal turnover is high, and urban principals are more likely to have a shorter tenure than principals in other settings (Fuller, Young, & Shoho, 2010; Levine, 2005; Young & Fuller, 2009). Administrator preparation programs that specifically focus on urban schools would give administrators the background in leading schools with the specific challenges listed in the six subdomains noted. Having targeted training may work to slow the turnover rate of principals in urban schools.

- The difference in administrator mean scores by school level and urbanicity, provide the opportunity to group leaders with leaders from the other subgroups for the purpose of on-going, job-embedded professional growth and development (Bowers, 2010, 2015; Bowers, Ni, & Esswein, 2018; Ni, Bowers, & Esswein, 2016). Collaboration between administrators from different contexts would help administrators gain knowledge and skills from each other on how to improve on various research-based leadership tasks (Bowers, 2010, 2015; Bowers et al., 2018; Ni et al., 2016). The high differences in scores between teachers and administrators on particular subdomains provides opportunities for districts to build capacity through on-going training and support can be given to principals around those particular domains. As CALL is a school improvement survey, this is one very practical use for this data for county and state offices.

References:

Freeman, M. (2020). *The Difference in Distributed Principal Leadership Practices Across Context According to the Comprehensive Assessment of Leadership for Learning*. [Doctoral Dissertation, Columbia University, Teachers College]. ProQuest Dissertations Publishing.

Halverson, R., & Kelley, C. (2017). *Mapping leadership: The tasks that matter for improving teaching and learning in schools*: John Wiley & Sons.

Suggested Citation:

Freeman, M. (2020). *The Differences in Principal Leadership Practices Due to Context*. CALL Research Group at Wisconsin Center for Education Products and Services.

Contact info: maf2284@tc.columbia.edu