

## **ABSTRACT**

GREBING, ERIC MICHAEL. An Investigation of the Impact of NC School Performance Grades on Teacher Perceptions and Turnover. (Under the direction of Dr. Stephen R. Porter).

This study used two-level sharp regression discontinuity models to evaluate the impact of the NC School Performance (A-F) Grades on teacher perceptions regarding support, autonomy, state assessments, their school as a good place to work and learn, and intent to remain teaching in their schools. The study exploited a natural experiment that isolated the impact of the grade label independent of other measures in the accountability model due to similarities in school performance near the thresholds between letter grades. Outcome data consisted of measures from the 2016 NC Teacher Working Conditions survey administered six months following the receipt of the A-F label for the 2014-15 school year. Examinations of the cutoff between failing (D or F) and passing (A, B, or C) yielded no significant discontinuities attributable to the performance label for any samples. Additional exploration of the B/C, C/D, and D/F cutoffs for elementary and middle schools also showed no significant discontinuities associated with the labels. The high school results showed a significant impact of the grade at the C/D cutoff for teacher perceptions of support and accuracy of state assessments. These results were robust to various model parameters and specifications.

© Copyright 2018 by Eric Michael Grebing

All Rights Reserved

An Investigation of the Impact of NC School Performance Grades on Teacher Perceptions and  
Turnover

by  
Eric Michael Grebing

A dissertation submitted to the Graduate Faculty of  
North Carolina State University  
in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy

Educational Research and Policy Analysis

Raleigh, North Carolina  
2018

APPROVED BY:

---

Dr. Stephen Porter  
Committee Chair

---

Dr. Alyshia Bowden

---

Dr. Anna Jacob

---

Dr. Paul Umbach

## DEDICATION

I am a first-generation doctoral student. Pursuing this degree would not have been possible without a strong network from people who love and care about me. I dedicate this dissertation to three people who have supported me consistently throughout this journey – my parents and my wife.

To Mom and Dad – from an early age, you instilled in me the value of education and service, surrounding me with constant opportunities to learn. Your unwavering support has helped me through personal and professional challenges as I have pursued my education. Over the years, I have learned a tremendous amount from you about perseverance and hard work – lessons that have helped me tremendously throughout doctoral coursework.

To Laurie – I would not be at this point of my educational journey if it had not been for your strong encouragement to pursue graduate school and setting a tangible example for me that completing a doctoral degree was possible. You consistently push me to grow personally and professionally. I look forward to presenting and publishing with you, especially now that we will no longer be “Doctor and Mister.”

## **BIOGRAPHY**

Eric Grebing began his educational journey in rural, public schools in Missouri and Illinois. He completed a bachelor's degree in electrical engineering and computer science from MIT in 2008, during which he completed coursework and student teaching to earn certification to teach high school physics in Massachusetts. Following a year of graduate engineering coursework and a teaching assistantship at MIT, Eric taught math and science at Warren Early College High School in northeastern North Carolina from 2009-2013. While teaching, Eric earned a master's in educational leadership from Concordia University-Portland. Eric then spent 2013-2016 as a data analyst and director of research and development at NC New Schools, an education non-profit. Since 2016, Eric has served as a research and evaluation specialist at the SERVE Center at UNC-Greensboro.

## ACKNOWLEDGMENTS

I want to thank all the members of the graduate education faculty and my dissertation committee for helping support me through this process. The experiences in this program have helped to shape my skills and perspectives as a researcher. In particular, Dr. Porter helped to push my thinking and was always available for questions and advice. His flexibility to meet with me through early morning calls and on weekends helped me to finish this work in balance with my professional commitments.

I would also like to acknowledge the professional experiences and the people involved that have allowed me to see various aspects of K-12 education and to learn from those with years of experience in the education system. I owe a debt of gratitude to the students, teachers, and administrators at Warren Early College High School, where I learned first-hand the great challenges and great rewards of working in education.

I would also like to acknowledge the talented educators I had the pleasure to work with at NC New Schools. The organization allowed me to learn from people with deep experiences as teachers, principals, and superintendents throughout North Carolina. Through this work, I gained a much deeper knowledge of K-12 public education in the state that undoubtedly helped me through the dissertation process.

Finally, I would like to acknowledge my colleagues at the SERVE Center. The past two and a half years have allowed me to gain deeper experiences in research and evaluation and have afforded me the opportunity to access the educational research profession. The daily conversations about our work and genuine care to help the field of education through research and evaluation have supported me throughout the final stages of my doctoral work.

## TABLE OF CONTENTS

LIST OF TABLES .....	vii
LIST OF FIGURES .....	ix
<b>Chapter 1: Introduction and Overview</b> .....	<b>1</b>
Background .....	1
North Carolina school accountability.....	1
A-F school grades.....	2
Pros and cons of A-F school grades.....	4
North Carolina School Performance Grades.....	6
Purpose of the Study .....	11
Theory of Action .....	12
<b>Chapter 2: Literature Review</b> .....	<b>18</b>
Impacts of High-Stakes Accountability Systems on Educators .....	19
Performance Information in School Accountability .....	21
Stigma and sanctions.....	22
Educator responses to school performance information in North Carolina.....	25
A-F School Grading Systems.....	28
Experimental impacts of A-F labels.....	29
Studies of A-F school grade policies.....	30
<i>Impacts on student outcomes</i> .....	38
Summary of A-F Research.....	42
Conclusion.....	43
<b>Chapter 3: Methods</b> .....	<b>45</b>
Analysis.....	48
Sample selection.....	48
Model specification.....	50
Model covariates.....	55
Treatment variable.....	56
Student demographics.....	59
Teacher characteristics.....	59
Staff turnover.....	60
Constructing Outcome Variables .....	61
Perceived support.....	62
Perceived autonomy.....	63
Perceived accuracy of state assessments.....	63
Good place to work and learn.....	64
Teacher turnover.....	65
Robustness Checks.....	66
Multiple comparisons.....	67
Sensitivity analysis.....	67
<b>Chapter 4: Results</b> .....	<b>69</b>
Research Question 1: Teacher Perceptions of Working Conditions in their Schools.....	71
Teacher perceptions of support.....	71
Perceived autonomy.....	80

Perceived accuracy of state assessments. ....	87
Perceptions of schools as good places to work and learn. ....	95
Research Question 2: Immediate Professional Plans .....	102
Intent to remain teaching at the same school next year. ....	102
Intent to leave education entirely. ....	111
Research Question 3: Actual Turnover .....	117
Sensitivity Analysis to Confirm Statistically Significant Results .....	125
<b>Chapter 5: Conclusions</b> .....	127
Connecting the Findings to the Theory of Action. ....	127
Suggested Areas for General Future Research on NC School Performance Grades .....	130
Explanations for and Mechanisms of Teacher Responses .....	132
Mechanisms to explain null findings. ....	132
Additional impact of the grades went undetected. ....	135
Mechanisms to Explain Significant Findings for the High School C/D Threshold .....	138
Recommendations and Policy Implications .....	139
Suggestions for future school report cards. ....	140
<b>References</b> .....	143
<b>Appendix A: Power Analysis</b> .....	155

## LIST OF TABLES

Table 1.1	States with A-F Grade Accountability Systems.....	3
Table 1.2	Conversion of School Performance Grade Score to Letter Grade.....	7
Table 1.3	Performance Measures Used to Calculate School Performance Grades .....	8
Table 2.1	Empirical Studies of A-F School Grades Literature Summary .....	32
Table 3.1	Descriptive Statistics for each Sample.....	56
Table 3.2	School Performance Grade Transition Matrix from 2013-14 to 2014-15 .....	57
Table 3.3	Number and Percentage of Missing and “Don’t Know” Responses.....	62
Table 3.4	Survey Items for Teacher Perception Outcomes by Construct.....	63
Table 3.5	Factor Loadings from Exploratory Factor Analysis on 2016 NC Teacher Working Conditions Survey .....	65
Table 3.6	Distribution of Teachers’ Immediate Professional Plans, 2016 NC Teacher Working Conditions Survey .....	66
Table 4.1	Perceptions of Support – Elementary School Sample .....	75
Table 4.2	Perceptions of Support – Middle School Sample .....	77
Table 4.3	Perceptions of Support – High School Sample.....	78
Table 4.4	<i>p</i> Values and Multiple Comparison Adjustment for High School Perceived Support .....	79
Table 4.5	Perceptions of Autonomy – Elementary School Sample.....	83
Table 4.6	Perceptions of Autonomy – Middle School Sample.....	85
Table 4.7	Perceptions of Autonomy – High School Sample .....	86
Table 4.8	Perceptions of State Assessment Accuracy – Elementary School Sample.....	90
Table 4.9	Perceptions of State Assessment Accuracy – Middle School Sample.....	92
Table 4.10	Perceptions of State Assessment Accuracy – High School Sample .....	94

Table 4.11	<i>p</i> Values and Multiple Comparison Adjustment for High School Perceived Accuracy of Assessments.....	95
Table 4.12	Good Place to Work and Learn – Elementary School Sample .....	98
Table 4.13	Good Place to Work and Learn – Middle School Sample .....	99
Table 4.14	Good Place to Work and Learn – High School Sample.....	101
Table 4.15	Intent to Stay Teaching at Same School – Elementary School Sample.....	104
Table 4.16	Intent to Stay Teaching at Same School – Middle School Sample.....	106
Table 4.17	<i>p</i> Values and Multiple Comparison Adjustment for Middle School Teacher Intent to Remain Teaching at Their Schools.....	106
Table 4.18	Intent to Stay Teaching at Same School – High School Sample .....	108
Table 4.19	Intent to Leave Education Entirely – Elementary School Sample.....	111
Table 4.20	Intent to Leave Education Entirely – Middle School Sample.....	112
Table 4.21	Intent to Leave Education Entirely – High School Sample .....	113
Table 4.22	Actual Teacher Turnover – Elementary School Sample.....	117
Table 4.23	Actual Teacher Turnover – Middle School Sample .....	119
Table 4.24	Actual Teacher Turnover – High School Sample .....	120
Table 4.25	Additional Specifications for Significant Results – High School Sample.....	122

## LIST OF FIGURES

Figure 1.1	Timeline of Events for the Study Outcomes Related to the NC Teacher Working Conditions Survey .....	6
Figure 1.2	Timeline of Events for the Study Teacher Turnover Outcome.....	6
Figure 1.3	Theory of Action.....	13
Figure 3.1	Timeline of Events for the Study Outcomes Related to the NC Teacher Working Conditions Survey .....	46
Figure 3.2	Timeline of Events for the Study Teacher Turnover Outcome.....	46
Figure 3.3	2014-15 NC School Performance Grade Distribution by Sample .....	50
Figure 4.1	Discrete Point Plots and Local Pass/Fail Regressions for Perceived Support .....	73
Figure 4.2	Local Regressions by Letter Grade for Elementary School Perceived Support .....	74
Figure 4.3	Local Regressions by Letter Grade for Middle School Perceived Support .....	76
Figure 4.4	Local Regressions by Letter Grade for High School Perceived Support .....	78
Figure 4.5	Discrete Point Plots and Local Pass/Fail Regressions for Perceived Autonomy.....	82
Figure 4.6	Local Regressions by Letter Grade for Elementary School Perceived Autonomy.....	83
Figure 4.7	Local Regressions by Letter Grade for Middle School Perceived Autonomy.....	84
Figure 4.8	Local Regressions by Letter Grade for High School Perceived Autonomy.....	86
Figure 4.9	Discrete Point Plots and Local Pass/Fail Regressions for Perceptions of State Assessment Accuracy .....	88
Figure 4.10	Local Regressions by Letter Grade for Elementary School Perceptions of State Assessment Accuracy .....	89

Figure 4.11	Local Regressions by Letter Grade for Middle School Perceptions of State Assessment Accuracy .....	91
Figure 4.12	Local Regressions by Letter Grade for High School Perceptions of State Assessment Accuracy .....	93
Figure 4.13	Discrete Point Plots and Local Pass/Fail Regressions for Perceptions of Schools as Good Places to Work and Learn .....	96
Figure 4.14	Local Regressions by Letter Grade for Elementary School Perceptions of Schools as Good Places to Work and Learn .....	97
Figure 4.15	Local Regressions by Letter Grade for Middle School Perceptions of Schools as Good Places to Work and Learn .....	98
Figure 4.16	Local Regressions by Letter Grade for High School Perceptions of Schools as Good Places to Work and Learn .....	100
Figure 4.17	Discrete Point Plots and Local Pass/Fail Regressions for Teacher Intent to Remain Teaching at Their Schools the Following Year.....	102
Figure 4.18	Local Regressions by Letter Grade for Elementary School Teacher Intent to Remain Teaching at Their Schools the Following Year.....	103
Figure 4.19	Local Regressions by Letter Grade for Middle School Teacher Intent to Remain Teaching at Their Schools the Following Year.....	105
Figure 4.20	Local Regressions by Letter Grade for High School Teacher Intent to Remain Teaching at Their Schools the Following Year.....	107
Figure 4.21	Discrete Point Plots and Local Pass/Fail Regressions for Teacher Intent to Leave Education Entirely .....	109
Figure 4.22	Local Regressions by Letter Grade for Elementary School Teacher Intent to Leave Education Entirely .....	110
Figure 4.23	Local Regressions by Letter Grade for Middle School Teacher Intent to Leave Education Entirely .....	111
Figure 4.24	Local Regressions by Letter Grade for High School Teacher Intent to Leave Education Entirely .....	113
Figure 4.25	Discrete Point Plots and Local Pass/Fail Regressions for Actual Teacher Turnover .....	115

Figure 4.26 Local Regressions by Letter Grade for Elementary School  
Teacher Turnover.....116

Figure 4.27 Local Regressions by Letter Grade for Middle School  
Teacher Turnover.....118

Figure 4.28 Local Regressions by Letter Grade for High School  
Teacher Turnover.....120

Figure 5.1 Theory of Action.....124

## CHAPTER 1

### Background

State accountability systems often attach labels to school performance. These labels, meant to hold schools accountable and improve educational outcomes for students, can influence educator perceptions and their decision to stay or leave a school. Given the importance of teachers in achieving educational outcomes, it is important to better understand how accountability policies affect them. A-F labels, in which states assign a letter grade of A, B, C, D, or F to schools on publicly-available school report cards, have grown in popularity throughout the United States over the past two decades (Howe & Murray, 2015). Although many states placed more ambiguous labels on schools prior to the A-F system, a “failing” label of D or F brings a nearly universal understanding of meaning and, with that, a greater ability to stigmatize schools (Figlio & Rouse, 2005). The following study explored the impact of the initial years of the A-F school grading policy in North Carolina, known as School Performance Grades, on teachers. Specifically, the study examined teacher perceptions, immediate professional plans, and subsequent teacher turnover by combining administrative data and responses from the biennial North Carolina Teacher Working Conditions Survey. This dissertation contributes unique evidence from the perspective of teachers to A-F school report card policy.

**North Carolina school accountability.** The ABCs of Public Education, an early program for high-stakes accountability in North Carolina, went into effect in 1996 for grades K-8 and for high schools beginning in 1998 (NC DPI, 2012). Under this program, North Carolina assigned performance labels (not letter grades) based on test scores to schools from the late 1990s through 2013 via the ABCs of Education policy (NC DPI, 2012). Under this precursor to the A-F School Performance Grades, the state calculated composite proficiency percentages on

End-of-Grade and End-of-Course tests and used the combined percentage and a measure of academic growth to assign labels to schools. These labels included (in order from lowest to highest) Low Performing, Priority School, School of Progress, School of Distinction, School of Excellence, and Honor School of Excellence (NC DPI, 2012). Beginning in 2002, schools also received labels related to Adequate Yearly Progress (AYP), a binary indicator of whether subgroups of students met a combination of annual performance targets on standardized tests. The ABCs policy and assessment of AYP ended in 2012 with the state's transition to Common Core State Standards and related testing (NC DPI, 2015). The end of this policy also coincided with the introduction of A-F School Performance Grades on North Carolina school report cards.

**A-F school grades.** Prior research has uncovered impacts of A-F school grades on student outcomes (Figlio & Rouse, 2005; Chiang, 2009; Rockoff & Turner, 2010; Winters & Cowen, 2012), public perceptions (Chingos, Henderson, & West, 2012; Figlio & Kenny, 2009; Charbonneau & Van Ryzin, 2012; Figlio & Lucas, 2004), and organizational changes (Chiang, 2009). As of September 2018, 15 states used A-F grades to label schools on annually-released school report cards. A-F systems help states to fulfill requirements under the Every Student Succeeds Act (ESSA) by providing a “summative determination” for each school based on required academic indicators (US Dept. of Education, 2017).

Table 1.1 summarizes the states with current A-F grading policies. Maine formerly used A-F school grades but abandoned the practice within the last five years. In addition, states such as Alabama (Crain, 2016a; Crain 2016b) and Arizona (Palmer, 2016) temporarily suspended using A-F grades, but returned to the practice as of 2018. New York City public schools also used an A-F system for their school report cards from 2007 to 2013 (Winters, 2016).

Understanding the national landscape helps to contextualize the North Carolina School

Performance Grade policy. The results from North Carolina could also inform policy throughout other states in understanding consequences of A-F school performance labels.

Table 1.1

*States with A-F Grade Accountability Systems*

State/City	First Year of A-F Grading	Last Year of A-F Grading	Sources
Alabama	2011-12	Current	(Crain, 2018)
Arizona	2009-10	Current	(Palmer, 2016)
Arkansas	2004-15	Current	(Arkansas News, 2015)
Florida	1998-99	Current	(Figlio & Rouse, 2005)
Georgia	2011-12	Current	(GA Gov. Office, 2017)
Indiana	2010-11	Current	(Elliott, 2013)
Louisiana	2010-11	Current	(Louisiana Dept. of Education, 2013)
Maine	2012-13	2014-15	(Maine Dept. of Education, 2015)
Mississippi	2010-11	Current	(Mississippi Center for Public Policy, 2012)
New Mexico	2010-11	Current	(NM Public Education Dept., 2015)
New York City*	2006-07	2012-13	(Winters, 2016)
North Carolina	2013-14	Current	(NC DPI, 2015)
Ohio	2012-13	Current	(Murray, 2013)
Oklahoma	2011-12	Current	(OK Dept. of Education, 2017)
Utah	2012-13	Current	(ExcelinEd, 2015)
Texas	2017-18	Current	(Tanner, 2016)
West Virginia	2014-15	Current	(Martriano & Green, 2016)

\*Although not a state system, the size of the district and style of school report card in NYC serves as another example of A-F grade assignment to a large sample of schools, similar to a state-wide system.

In almost all states and cities that have implemented school letter grades, including North Carolina, some form of high stakes accountability existed prior to assigning letter grades on school report cards. Some states made the initial shift to A-F grading by simply translating former discrete performance labels into a corresponding letter grade. For example, Mississippi translated the labels of “star,” “high performing,” “successful,” “academic watch,” and “low performing/failing,” to A, B, C, D, and F respectively (Mississippi Center for Public Policy, 2012). In a similar manner, Indiana adapted the labels of “exemplary,” “commendable,” “academic progress,” “academic watch,” and “academic probation” to the A-F labels (Elliott, 2013). In Florida, however, the introduction of the A+ Schools Initiative in 1999 implemented

high stakes accountability and school letter grades simultaneously (Figlio & Lucas, 2004).

Discussed in detail in the later section on North Carolina school accountability, the magnitude of the shift in North Carolina lies between the extremes of a simple transfer of labels and the introduction of accountability; the School Performance Grades labeled school performance based on a new battery of tests and standards the school year after North Carolina adopted the Common Core State Standards in 2012-13 (NC DPI, 2015).

**Pros and cons of A-F school grades.** School-level A-F performance grades (i.e., assigning schools one of five performance categories of A, B, C, D, or F) aim to provide easily understood information about school quality to a variety of public stakeholders (Coe & Brunet, 2006). An A-F label provides informational accessibility in that any child or adult who has experienced schooling likely encountered letter grades throughout his or her education. This style of grading has been used beyond the classroom for decades, offering a way to assess the status and progress of a variety of institutions from restaurant sanitation to infrastructure quality (Coe & Brunet, 2006).

Despite easy interpretability, A-F grading systems applied to institutions have undergone critiques from researchers and policymakers (Adams et al., 2013; Howe & Murray, 2015). Lower grades are associated with differences in community perceptions of schools (Chingos, Henderson, & West, 2012; Favero & Meier, 2013; Jacobsen, Snyder, & Saultz, 2014), school environments and resources (Chiang, 2009; Rouse et al., 2013), and teachers' satisfaction with their working environments (Ladd & Linderholm, 2008; Favero & Meier, 2013).

Additional research on student outcomes offers an alternative outlook about the impact of A-F school grades. Proponents of A-F policies cite the simplicity of the measures as an asset in “transparency” (Howe & Murray, 2015) and have empirical evidence of improved student

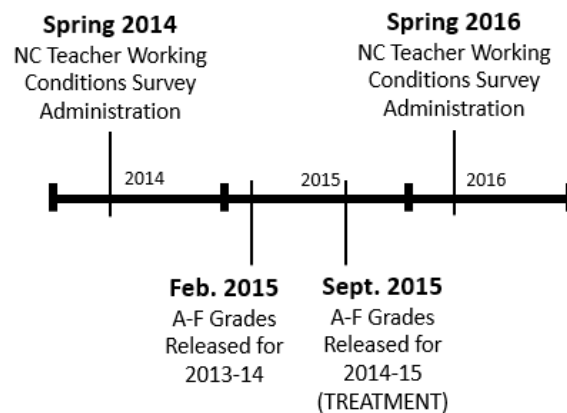
outcomes in schools following the receipt of low grades, particularly F's (Figlio & Rouse, 2005; Chiang, 2009; Rockoff & Turner, 2010; Winters & Cowen, 2012).

Experimental evidence has demonstrated that people react differently to A-F labels than to other formats of school performance data (Jacobsen, Snyder, & Saultz, 2014; Ladd & Linderholm, 2008). Even when numerical performance data accompany A-F letter grades, people are unlikely to look beyond the letter grade when forming judgements about performance and quality (Jacobsen, Snyder, & Saultz, 2014; Olsen, 2013). In experimental settings, people form more polarizing perceptions of the quality of schools (Jacobsen, Snyder, & Saultz, 2014) and perceptions of student behavior (Ladd & Linderholm, 2008) when schools are labeled with an A-F grade.

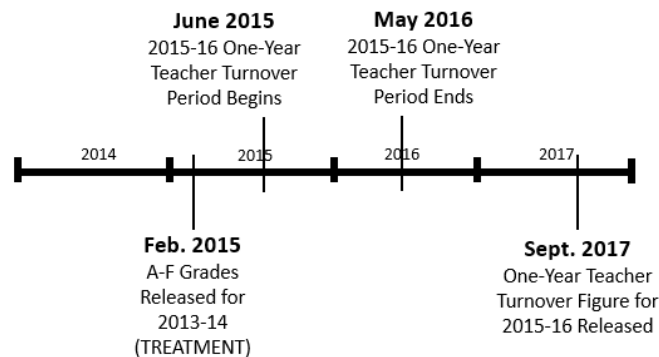
North Carolina provides a unique opportunity to investigate how A-F school labels impact teacher perceptions. The advent of School Performance Grades for the 2013-14 school year built upon prior school labeling, attaching a single A-F letter grade to each school based on performance and student growth. The A-F system introduction, however, identified a much larger proportion of schools as “failing” than the previous ABCs system. In 2012, only 15 of over 2,000 schools receiving a rating were labeled as “Low Performing,” with an additional 160 schools labeled as “Priority Schools.” By contrast, in 2014, 707 schools (29.1% of schools that received a letter grade), received a School Performance Grade of D or F (NC DPI, 2015).

NC General Statute §115C-105.37 also changed the definition of a “low performing school” to those who receive a D or F school performance grade and a growth score of “met expected growth” or “not met expected growth” (NC DPI, 2015). For the first round of grades released in February 2015, 621 North Carolina schools, or 25.6% of schools receiving a letter grade, were designated as low performing under this definition (NC DPI, 2015). The increase in

the number of low performing schools had a potential stigmatizing effect. The introduction of these labels serves as a natural experiment, because many schools that were not previously identified as failing now had a D or F label attached to them. Figure 1.1 describes the timing of events for outcomes related to the Teacher Working Conditions survey. Figure 1.2 shows the timing of events for the teacher turnover outcome.



*Figure 1.1.* Timeline of events for the study outcomes related to the NC Teacher Working Conditions survey.



*Figure 1.2.* Timeline of events for the study teacher turnover outcome.

**North Carolina School Performance Grades.** Passed in 2013 by the NC General Assembly, G.S. 115C-83.15(b) articulates the components used to calculate School Performance Grades. As of September 2018, the calculation of grades still follows the initial guidelines. The

overall School Performance Grade score consists of two parts: the achievement score and the EVAAS growth score. The school achievement score combines proficiency percentages of state end-of-grade and end-of-course tests, graduation rates, the percentage of students scoring a 17 or higher on the ACT, the percentage of Career Technical Education-concentrating students earning Silver level or higher on the WorkKeys exam, and the percentage of students successfully completing Math III or higher. Each school also receives an EVAAS value-added growth score scaled from 50 to 100 based on how students perform relative to predicted scores. The overall School Performance Grade score is a weighted average of 80% achievement and 20% growth. The School Performance Grade is then translated directly to a letter grade on a 15-point scale. Table 1.2 contains the conversion values of the School Performance Grade score to letter grades.

Table 1.2

*Conversion of School Performance Grade Score to Letter Grade*

<u>School Performance Grade</u>	<u>School Performance Grade Score Range</u>
A	85-100
B	70-84
C	55-69
D	40-54
F	0-39

The state uses different measures to create the composite score depending on whether they describe elementary, middle, or high school grade levels. For high schools, standardized test scores accounted for only a portion of the academic achievement composite with the measures for graduation rate and third-year math course completion (math course rigor) raising the performance composite. The grades of elementary and middle schools, however, are solely based on test scores for the academic achievement composite. Table 1.3 details the measures used to determine the grade for each school type. Due to these differences, the study examined impacts separately for elementary, middle, and high school samples. I determined the sample designation

to each school based solely on the performance measures listed in Table 1.3. I eliminated schools from the study with combinations of performance metrics that did not align to the definitions for elementary, middle, and high schools. Chapter 3 contains a more detailed description of the process for sample selection.

Table 1.3  
*Performance Measures Used to Calculate School Performance Grades*

Grade Levels	Variable	Description, from §115C-83.15(b)
Elementary and Middle Schools	Math End-of-Grade Test Percent Proficient	One point for each percent of students who score at or above proficient on annual assessments for mathematics in grades three through eight.
	Reading End-of-Grade Tests Percent Proficient	One point for each percent of students who score at or above proficient on annual assessments for reading in grades three through eight.
	Science End-of-Grade Tests Percent Proficient	One point for each percent of students who score at or above proficient on annual assessments for science in grades five and eight.
Middle and High Schools	Math I End-of-Course Test Percent Proficient	One point for each percent of students who score at or above proficient on the Algebra I or Integrated Math I end-of-course test.
	English II End-of-Course Test Percent Proficient	One point for each percent of students who score at or above proficient on the English II end-of-course test.
High Schools	Biology End-of-Course Test Percent Proficient	One point for each percent of students who score at or above proficient on the Biology end-of-course test.
	Four-Year High School Graduation Rate	One point for each percent of students who graduate within four years of entering high school.
	Math Course Rigor	One point for each percent of students who complete Algebra II or Integrated Math III with a passing grade.
	Percent of Students Scoring 17 or Higher on ACT	One point for each percent of students who achieve the minimum score required for admission into a constituent institution of The University of North Carolina on a nationally normed test of college readiness.
	Percentage Earning Silver Level or Higher on WorkKeys Exam	One point for each percent of students enrolled in Career and Technical Education courses who meet the standard when scoring at Silver, Gold, or Platinum levels on a nationally normed test of workplace readiness.

At the time of this study, the letter grades alone did not carry direct sanctions such as threat of immediate takeover or losing students from schools through a voucher program, as was done in other states such as Florida. However, the 2014-15 SPG signified the first year that the letter grade was tied to “low performing” status. The 2013-14 grade did not have bearing on low performing status; the criteria from the previous ABCs accountability policy was still in effect. According to G. S. 115C-83.15, “low-performing schools are those that receive a School Performance Grade of D or F and a school growth score of “met expected growth” or “not met expected growth.” The actions required of schools receiving a low-performing designation could lead to some additional stigma through having to notify parents and submit publicly-facing plans to address low performance. Superintendents also had to determine whether to retain, develop, transfer, or dismiss the principal of the school. This additional consequence of the 2014-15 grades provided further possibility of an impact of the grades on teachers.

**NC Teacher Working Conditions Survey.** North Carolina surveys all teachers every two years using the NC Teacher Working Conditions Survey (Maddock, 2009). With an 86% response rate from certified staff across the state in 2016 (New Teacher Center, 2018), the survey results offer a biennial snapshot of teacher perceptions for nearly every school in the state. The combination of letter grades with the survey allows the exploration of how the A-F labels impact teacher perceptions.

The survey measures eight constructs with subscales. These constructs include *Time*, *Facilities and Resources*, *Community Support and Involvement*, *Managing Student Conduct*, *Teacher Leadership*, *School Leadership*, *Professional Development*, and *Instructional Practices and Support*. Published most recently for the 2014 survey, these subscales had high reliability with Cronbach’s Alpha values ranging from .86 to .96 for the eight subscales (New Teacher

Center, 2014). No subscale in its full form conceptually matched the expected impact of labeling a school with an A-F grade, so I selected eight relevant items from the survey, grouped into two three-item scales and two single-item outcomes. Chapter 3 contains more detail about the items selected for the analysis.

### **Statement of the Problem and Significance**

Work in evaluating A-F grade policies in other locations, predominantly Florida (Figlio & Lucas, 2004; Figlio & Rouse, 2005; Chiang, 2009; Figlio & Kenny, 2009; Feng, Figlio, & Sass, 2010; Chingos, Henderson, & West, 2012; Rouse et al., 2013) and New York City (Rockoff & Turner, 2010; Winters & Cowen, 2012; Charbonneau & Van Ryzin, 2012; Favero & Meier, 2013; Jacobsen, Saultz, & Snyder, 2013; Dizon-Ross, 2014; Winters, 2016), focused on student outcomes and community perceptions, with less emphasis on teacher perceptions. Favero and Meier (2013) connected A-F grades with teacher surveys, but the study only looked at correlations of perceptions with measures rather than measuring changes in teacher perceptions after the school received a grade. In addition, some studies have explored the impact of A-F school grading on teacher turnover (Feng, Figlio, & Sass, 2010; Dizon-Ross, 2014). Despite the ubiquity of A-F school report card labels throughout the United States, however, a clear understanding of how these policies impact teachers does not yet exist.

In addition, little research currently exists on the impacts of the School Performance Grade policy in North Carolina (Pierson et al., 2015; Smith & Imig, 2017; New Teacher Center, 2018). Pierson et al. (2015) evaluated the way in which the state calculated grades, citing issues with a high correlation of grades with poverty and too little emphasis on student growth. In addition, Smith and Imig (2017) surveyed principals to understand their perceptions of the new A-F policy. Most recently, a 2018 report from the New Teacher Center noted *correlations*

between School Performance Grades and all major scales on the Teacher Working Conditions survey. The correlation noted in the report reflects the underlying performance composite from schools, not the impact of the label placed on the school. As such, no research to date has studied how the A-F grade labels assigned to schools in North Carolina impact teachers' perceptions or decisions to remain at their schools. Given the ubiquity of A-F systems throughout the country, understanding this topic makes a potentially substantial contribution to the literature on accountability and teacher perceptions.

### **Purpose of the Study**

Teacher perceptions of their schools matter for student achievement and teacher turnover. Leithwood and McAdie (2010) posited that teachers' perceptions, or internal states, serve as the immediate causes of what teachers do on the job. Following this argument, the authors found that positive school environments have positive effects on teachers, which in turn, enhance their ability to educate their students. Additionally, Sabin (2015) found a positive relationship between teacher morale and student academic growth in North Carolina. Adverse working conditions were also associated with higher teacher turnover in California (Loeb, Darling-Hammond, & Luczak, 2005).

The following study aimed to understand the impact of the placement of a failing School Performance Grade label on teacher perceptions of their schools. The results provide evidence for policymakers in North Carolina and other states with A-F school grading policies to understand whether the grades assigned to schools have unintended consequences on teacher perceptions and teacher turnover.

## **Research Questions and Hypotheses**

This dissertation addressed the following questions:

- 1) Does the School Performance Grade impact teachers' perceptions of a) support, b) autonomy, c) accuracy of state assessments, and d) their schools as good places to work and learn?
- 2) Does the School Performance Grade impact teachers' immediate professional plans in terms of a) intending to remain teaching at their same school the following year and b) intending to leave education entirely?
- 3) Does the School Performance Grade impact subsequent teacher turnover?

I hypothesized that the receipt of a D or F grade for 2014-15 would have a negative impact on teacher perceptions of their schools in the domains addressed in the first research question on the 2016 NC Teacher Working Conditions Survey. Additionally, I hypothesized that the receipt of a D or F grade for 2014-15 would decrease the rate of teachers planning to remain teaching in their schools and increase the rate of teachers planning to leave education entirely as indicated in the 2016 NC Teacher Working Conditions Survey. Finally, I hypothesized that the receipt of a D or F grade for 2013-14 would increase school-level teacher turnover in the 2015-16 school year, reported as the one-year turnover figure in the 2017 personnel data file.

## **Theory of Action**

The following theory of action guided the investigation, summarized in Figure 1.3. The letter grade label assigned to the school can impact the community's perceived quality of the environment (Charbonneau & Van Ryzin, 2012). The change in perceived quality can also influence the level of support for the school (Chingos, Henderson, & West, 2012). Teachers bear

responsibility for the performance of their students and may feel diminished support from parents and the community if the school receives a failing grade.

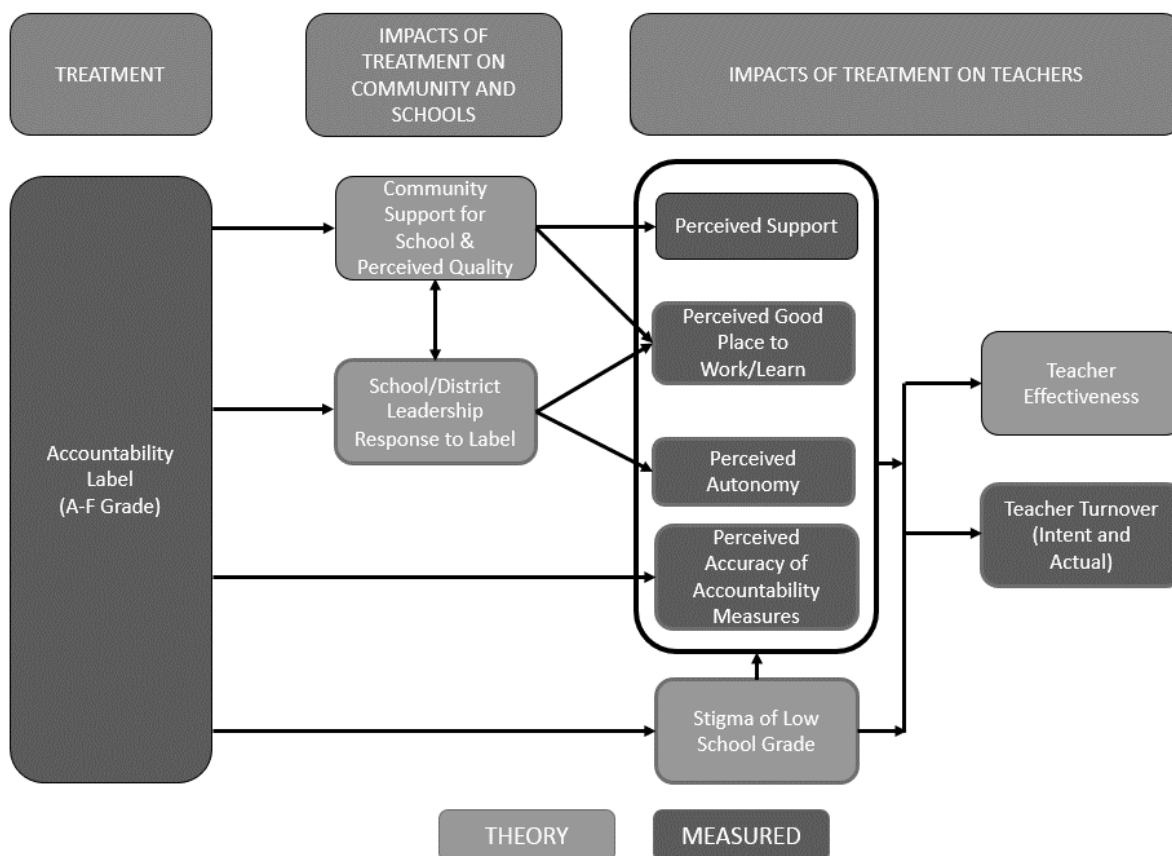


Figure 1.3. Theory of action.

Additionally, the letter grade label can affect school and district leadership through stigmatizing effects or through the threat of sanctions due to the accountability label (Figlio & Rouse, 2005; Chiang, 2009; Rouse et al., 2013). Together, these factors can influence the actions of community members toward the school (Figlio & Kenny, 2009) and the level of autonomy provided to schools (Reed et al., 2001). Due to pressure to improve, leadership may demonstrate diminished trust in teachers and lead to a loss of perceived autonomy for teachers. The combination of the stigma from the labels and the diminished support of community members and school and district leaders may also lead to a perceived worsening of overall school climate.

The grade assigned to a school, particularly a failing grade, could also directly impact teachers' belief that assessments accurately gauge student understanding. A survey with North Carolina principals showed a lack of trust in the reliability of accountability measures when coupled with a low grade (Smith & Imig, 2017). It is likely to impact teachers in a similar manner. Pierson et al. (2015) also noted that the letter grades strongly correlate with poverty. The heavy emphasis on areas outside of a school's control could lead teachers to negatively view the accuracy of the accountability measures.

The working conditions perceived by teachers support their ability to be effective in their roles (Leithwood & McAdie, 2010). Although the proposed study will not measure teacher effectiveness, this component directly influences the grade that a school will receive in future years. Teacher perceptions of their working environments can also impact teacher turnover via four mechanisms. First, teachers may choose to leave because of the stigma of teaching in a low performing school (Chiang, 2009). Teachers may also elect to leave a school due to perceived deterioration of working conditions (Johnson, Kraft, & Papay, 2012). School leadership may react to a low grade by trying to remove teachers deemed to not be effective. Finally, an opposite phenomenon could also occur if teachers choose to stay in their schools to contribute to future improvement (Dizon-Ross, 2014).

### **Research Design**

The study utilized a cross-sectional data set of North Carolina schools from the 2016 NC Teacher Working Conditions Survey and corresponding school-level student demographic, student performance, and teacher demographic data from 2013-14, the last school year prior to treatment. The teacher-level data from the survey allowed for a two-level model to assess the first two research questions; the third research question, focused on school-level teacher

turnover, used a single-level model. Due to the sharp cutoffs of grades that map to a continuous score, I used a regression discontinuity design (RD) with the School Performance Grade Score as the running variable and the A-F grade that a school received in 2014-15 as the treatment for RQ1 and RQ2 and the grade from 2013-14 to determine the treatment for RQ3. For RQ1 and RQ2, I used school-level NC Teacher Working Conditions Survey results from 2014 as “pretest” covariates. In all models, I included relevant school-level student and staff characteristics as covariates to increase statistical power.

As noted previously in this chapter, the performance measures used to calculate the School Performance Grade score and letter grade depend on the type of school. These different school environments could respond differently to the grade label. Thus, I ran separate analyses for elementary, middle, and high schools. I conducted a “pass” vs. “fail” analysis for all schools at the C/D grade cutoff for all three school-type samples. This cutoff for “passing” also contains inherent meaning in that any school with a School Performance Grade score of 55 or higher (the minimum value for a C) is not eligible for “low performing” status in that year. Although the individual grade thresholds had lower statistical power to detect an effect, I also assessed the impact of a B vs. C, C vs. D, and D vs. F for each of the three samples for cases with a sufficient sample size. Appendix A contains the power calculations used to determine the feasibility of this analysis.

### **Limitations**

The limitations of the proposed study fall into three major categories: the sample, the North Carolina Teacher Working Conditions Survey instrument, and the inference of causality. The data used for the study only apply to non-schools of choice in North Carolina. The results are not generalizable to charter schools, early college high schools or to schools outside of North

Carolina. In the 2014-15 school year, 2,589 schools recorded student performance data (NC DPI, 2015). Of these schools, 2,446 (94%) received a valid School Performance Grade. The samples used in the study with complete data for all observations and meeting other criteria for selection described in Chapter 3 included 1,968 public schools, accounting for 76% of all North Carolina public schools and 80% of all schools that received a letter grade. In addition, the measures used to calculate School Performance Grades differed by the grade levels served. Thus, each analysis took place independently for elementary, middle, and high schools, instead of collectively for all schools in the state. The reduction in the sample size for individual analyses decreased statistical power, not allowing for analysis at each grade threshold for all samples.

Second, the study measured teacher perceptions using items from the North Carolina Teacher Working Conditions Survey. As such, the study was limited by relying on self-reported outcomes only collected every two years. The timing of survey administration also did not occur quickly after the receipt of a school grade; the 2016 survey window took place seven months after the latest grades were released to the public. Another instrument-related limitation dealt with the measurement of the proposed constructs in the first research question about teacher perceptions. Although the constructs represented in the research questions are rooted in prior research as conceptually relevant to the School Performance Grade policy, the items do not explicitly measure teacher responses to the grade. The survey also did not collect any identifiable teacher data and thus, the study could not connect teacher responses to the same individuals from year to year. In the multilevel context, this eliminated the ability to include Level 1 covariates or the creation of a teacher-level panel data set.

Finally, the study attempted to infer causality in the absence of a controlled experiment. Given the complexity of school environments, it was challenging to isolate the School

Performance Grade as the only explanation for any changes in teacher perceptions. The study design mitigated the threats to inference by using many available administrative data variables as covariates. In addition, the RD offered a strong quasi-experimental approach by closely mimicking random assignment around the cutoff points of grade assignment (Lee & Lemieux, 2010).

## CHAPTER 2

School labels have become synonymous with school report cards and accountability. From labels such as “low performing,” “met AYP,” and A-F letter grades, prior research has established that these labels can influence perceptions and actions (Clotfelter et al., 2004; Ladd & Linderholm, 2008). School accountability has evolved in the United States over several decades. During the 1980s and 1990s, states began to formally measure student performance and provide data to the public (Figlio & Loeb, 2011). During the 1990s, state leaders positioned North Carolina as an early champion of high-stakes public school accountability. Along with Texas and a few other early adopting states, North Carolina tested its students and disseminated information about school performance to the public via school report cards years before federal mandates were in place (Figlio & Loeb, 2011).

The 2001 passage of No Child Left Behind (NCLB) required public dissemination of school accountability data for all states, particularly standardized test scores and graduation rates (Grissom, Nicholson-Crotty, & Harrington, 2014). Making measures of student achievement available to the public increased transparency but raised the level of scrutiny for public schools (Murillo and Flores, 2002; Mintrop and Trujillo, 2005). The dissemination of school performance measures to the public via school report cards can inform parental and community judgments about their local schools (Charbonneau & Van Ryzin, 2012).

The data shared via school report cards changed to include different measures and representations of school performance throughout the first decade of the 2000s to the present. Although the metrics and format of the data presented differ by state, all 50 states, the District of Columbia, and many large urban school districts independently publish their school performance data to the public on an annual basis.

The following literature review attempts to explain why A-F grades may impact teacher perceptions in North Carolina based on prior scholarship. The review begins with an overview of the potential impacts of high-stakes accountability systems on educators. Following this introduction, the review explores the role of public dissemination of school performance information in shaping opinions and decisions related to public school quality. Finally, the review explores the case of A-F school grading systems through impacts in controlled experiments and evidence from real systems involving public and parental perceptions, school leaders, student achievement, and teacher perceptions.

### **Impacts of High-Stakes Accountability Systems on Educators**

Accountability policies can have a direct impact on educators through higher stress and expectations and can impact how educators perceive their working environments. A-F grading systems are a special case of accountability systems that contain less ambiguous labels than numeric composite scores or categorical indicators that do not have a universal meaning. The following section provides detail about the impacts of general high-stakes accountability systems on educators.

Noting an absence of teacher voice in early stages of North Carolina high-stakes school accountability, Jones et al. (1999) conducted a survey of teachers in a sample of 16 elementary schools in North Carolina following the initial implementation of the ABCs accountability system. According to the research team's survey results, 77% of teachers reported lower morale, 76% indicated that their jobs were more stressful and 76% also stated they did not believe the accountability program would improve the quality of their schools.

Lyons and Algozzine (2006) surveyed North Carolina principals about their perceptions of the role of accountability in their schools early into the NCLB era. The authors showed that

principals unfavorably viewed testing requirements, sanctions, and the labels applied to schools based on student achievement.

Finnigan and Gross (2007) conducted a mixed methods study in 10 low-performing Chicago elementary schools to understand teacher motivation in the context of NCLB. The study involved interviews, focus groups, and surveys with teachers. The authors found that the low-performing status of a school led to low teacher morale, despite improvements in student achievement seen during the NCLB-era accountability policies. They conjectured that low morale can "undercut" the sustainability of change in teaching practices despite short-term performance gains stating, "Decreased morale threatened to reverse any increases in effort and changes in practice in schools that remained on probation" (p. 624). The study provided evidence that accountability policies, especially when they highlight poor performance, affect teachers and can counteract initial positive student gains.

Grissom, Nicholson-Crotty, and Harrington (2014) used a difference-in-differences approach to investigate impacts of initial implementation of NCLB on states with no prior accountability system in place versus those with an existing high-stakes system. The authors used data from the nationally-representative Schools and Staffing Survey, an instrument used to explore variables about school personnel at a national level. Using data from four administrations of the survey from 1994-2008, the authors found that no substantial differences in teacher perceptions were attributed to having a pre-NCLB high-stakes accountability system in place. In other words, the study found no adverse effects of accountability on teachers.

Sun, Saultz, and Ye (2016) used data from the nationally representative 1993-2009 Schools and Staffing Survey to study the impact of NCLB on teacher turnover. Using a differences-in-differences approach, the authors analyzed differences in attrition rates between a

treatment group of states that did not have a high-stakes accountability system in place prior to the 2002 adoption of NCLB and a comparison group of states that had a system prior to NCLB. They found that the introduction of NCLB increased involuntary turnover, more than doubling the likelihood of a teacher being involuntarily moved to another school.

In summary, the literature on how accountability systems directly impact teachers indicates mixed results. Descriptive studies indicated lower morale, as descriptive survey data early in North Carolina accountability systems showed lowered teacher morale, higher stress, and belief that the system would not improve schools (Jones et al., 1999). High-stakes accountability can also negatively impact principal perceptions of the state system (Lyons & Algozzine, 2006). Application of a low-performing label can also lead to lower teacher morale (Finnegan & Gross, 2007), potentially undermining improvements in student achievement that accountability policies aim to impact. Two relevant studies utilized nationally representative data from the Schools and Staffing Survey to explore impacts of NCLB on teachers. A more rigorous method by Grissom, Nicholson-Crotty, and Harrington (2014) showed no impact of the introduction of a high-stakes accountability policy on teacher morale and perceptions of their schools. The introduction of NCLB also increased involuntary teacher turnover, particularly in teachers moving to other schools (Sun, Saultz, & Ye, 2016).

### **Performance Information in School Accountability**

Labels placed on schools and on individual teachers and students are common elements of high-stakes accountability systems. Labeling schools based on their performance can influence educators and the public, evoking a response that extends beyond school walls. Prior literature suggests that people pay attention to, form opinions of, and make decisions in response

to public performance information about schools. The subsequent sections review prior studies that explore how people respond to school performance information.

**Stigma and sanctions.** There are two mechanisms by which school performance information can impact responses – those associated with stigma and others associated with sanctions (Figlio & Rouse, 2005). When a school is labeled with an F grade or as “low performing,” the stigma associated with the label can cause changes in teachers’ judgments about classroom behavior (Ladd & Linderholm, 2008). In addition, when such labels are coupled with sanctions, or consequences of underperformance, schools may respond in other ways that impact work climate and student achievement. Not surprisingly, sanctions in the form of additional oversight and restricted autonomy can cause changes to school environments. The stigma associated with a label can also impact school environments, making it worthy of attention in research (Murillo & Flores, 2002; Goldhaber & Hannaway, 2004; Figlio & Rouse, 2013; Jacobsen, Snyder, & Saultz, 2014; Bowen & Trivitt, 2014; King-Sears & Baker, 2014).

Previous policy studies (Figlio & Rouse, 2005; Saw et al., 2017) have attempted to disentangle the stigma of a label from accompanying accountability sanctions. Figlio and Rouse (2005) used the Florida “Critically Low Performing List,” a label with no sanctions attached, and compared it to the subsequent A-F grading label, for which receiving an F grade had a sanction of making students in a failing school eligible to receive a voucher to attend a private school. Seeing no significant difference in student performance for the schools labeled as “critically low performing” and those labeled with an “F,” the authors asserted that stigma on its own serves as a motivating factor for school changes.

In a study of the public context, Mintrop and Trujillo (2005) explored early accountability systems in place prior to NCLB. Their analysis highlighted that schools receiving

negative performance labels such as “low performing” received more intense scrutiny from evaluation teams, accountability requirements such as writing school improvement plans, and “mild public stigma” due to the label itself (p. 5). At the school level, Murillo and Flores (2002) characterized the stigma associated with low performing school labels as “reform by shame.” Although the authors described that the labels can catalyze positive change, their study of the North Carolina school accountability from 1995-2000 qualitatively captured the sentiment of educators as “diminishing” their role and value (p. 94). This finding accentuated the importance of understanding teacher perceptions in the context of school accountability labels.

Saw et al. (2017) explored the effects of two accountability labels in Michigan beginning in 2010. The first label of “persistently low achieving,” identified the lowest 5% of schools and has both public release of the list and accountability requirements from the state. The “watch list,” representing the 6<sup>th</sup> to 20<sup>th</sup> percentile of schools had no accompanying sanctions or media coverage associated with the label. Subsequent student performance in the “persistently low achieving schools” improved, but there was no improvement in the “watch list” schools. The effect of the “persistently low achieving” label and its associated consequences, paired with the null effects of the “watch list” label, suggest a strengthened effect when sanctions accompany stigma. The study also provides evidence that different labels placed on schools can elicit different responses, especially when certain negative labels receive greater public attention; the “watch list” moniker is an example of an ambiguous label that likely would not have much of an impact without publicity or sanctions. The literature demonstrates that both stigma and direct sanctions associated with performance labels can impact responses.

**Responses of non-educators to performance information.** High-stakes accountability can impact parents and the public through the dissemination about information about schools.

These measures may have an indirect impact on teachers through changed perceptions and behaviors of parents and community members toward education. The information gain for people outside of schools is potentially higher for those outside of education because, unlike educators, an annual school report represents a much larger proportion of the total information they receive about their local schools. In the theory of action for the study, the response of non-educators is most likely to contribute to teacher perceptions of support from parents and the community.

In his book addressing the dynamics of performance management, Moynihan (2008) noted that performance data play an integral part in allowing people to judge their public institutions with greater accuracy and improve decision making. Researchers have also explored the more specific case of public perceptions from public education accountability systems and the labels associated with them, as they aim to empower citizens with the information to hold schools accountable (Chingos, Henderson, & West, 2012; Jacobsen, Snyder, & Saultz, 2014). These impacts have been directly observed through perception surveys and indirectly through market impacts such as fundraising and housing prices.

Assessing the potential impacts of accountability systems and labels requires understanding how people utilize school performance information. It is difficult to ascertain how people directly respond to school performance information. The choices that parents make about where to send their students to school provides one opportunity to observe how people utilize performance data.

Hastings and Weinstein (2008) found that parents are more likely to choose higher-performing schools when presented with data about school performance. The study exploited two experiments involving disbursement of performance information to parents in Charlotte-

Mecklenburg Schools in North Carolina. First, a natural experiment occurred in 2004 when parents in schools sanctioned because of failure to meet AYP received information on the performance of their child's school and all other schools in the district. The authors also conducted a field experiment in 2006 in which a set of treatment parents received a one-page table of test scores and school choice options. Hastings and Weinstein showed that public data about school performance inform parental judgments about schools for school selection.

Although a non-surprising result, the authors found that parents choose higher performing schools when presented data about school performance in an easily accessible manner. The paper also illustrated that performance information coupled with good choice options can aid increases in academic achievement. Although prior to the assignment of A-F grades to schools in North Carolina, this study demonstrated that school performance information impacted North Carolina parental perceptions.

Friesen et al. (2012) also looked at school choice based on information in school report cards using a sample of students in British Columbia, Canada. The researchers used a difference-in-differences methodology that explored student mobility before and after the public availability of test score data. The study found a higher likelihood of students leaving their school when unfavorable school performance measures were released. Like Hastings and Weinstein (2008), this study strengthened the argument that parents utilize and respond to public information about schools, particularly when information negatively portrays a school. Together, these studies demonstrate that parents respond to performance information and make different choices about where to educate their students as a result.

**Educator responses to school performance information in North Carolina.** In 1996, North Carolina began to implement the ABCs accountability system for K-8 schools, which

incorporated high-stakes testing and labeled schools as “low performing” for not meeting certain targets. In this literature, schools labeled as “low performing” in North Carolina had higher rates of teacher turnover (Clotfelter et al., 2004). Heissel and Ladd (2016) discovered that teachers in schools receiving services related to school turnaround sanctions had an increase in professional development hours per week, and a decrease in independent planning time, but no change in measures of school climate. Gershenson (2016), however, found a *decrease* in teacher absences driven by a schools failure to meet AYP. In this case the application of a negative label had a positive impact on teacher attendance. The following section provides greater detail about these studies.

The Clotfelter et al. (2004) study explored teacher turnover in North Carolina by studying differences between two cohorts of teachers – those who began teaching in 1994-95 prior to a statewide system of high stakes accountability and those who began teaching in 1996-97 in the first years of the accountability system. They compared the schools with low student performance (< 50% proficiency) and did not meet growth and were thus labeled as “low performing” versus similarly low performing schools whose growth allowed them to not receive the label. Thus, the study controlled for student performance and isolated the label. After controlling for the differences among cohorts and school and teacher demographic characteristics, low-performing schools had significantly higher turnover than their higher-performing counterparts. They also wrote that the system “had an adverse effect on schools serving low-performing students by making them even less able to retain teachers than otherwise would have been the case” (p. 252). The study fits within a body of literature that impacts on teachers ultimately impact students, especially when policies drive teacher turnover.

Before the implementation of the A-F system, the “low performing” label represented the lowest categorization of school performance. Application of the “low performing” designation affected the ways in which teachers spent their time. Heissel and Ladd (2016) used the sample of North Carolina schools identified with the "low performing" label for turnaround actions under Race to the Top. Utilizing a regression discontinuity for schools near the cutoff for receiving the sanctions and services associated with the "low performing" distinction, Heissel and Ladd combined NC Teacher Working Conditions Survey data with teacher and principal turnover percentages to assess the impact of the turnaround program. The study generated a “school climate” scale that combined selected items from the *School Leadership*, *Instructional Practices*, and *Professional Development* survey subscales. The authors discovered that teachers in schools receiving services related to school turnaround sanctions had an increase in professional development hours per week, a decrease in independent planning time, but no change in measures of school climate. The study adds to the evidence base that sanctions associated with school accountability can change practices within schools that receive them.

Negative accountability labels can also impact teacher behavior in a positive way by reducing absences. Gershenson (2016) explored changes in North Carolina teacher absences in the first school year after schools failed meet Adequate Yearly Progress (AYP) under NCLB, making the argument that absences can serve as a proxy for teacher productivity and effort and offer a useful evaluation of the impact on teacher behavior. Gershenson obtained teacher-level data for Kindergarten through 5<sup>th</sup> grade teachers from 1997-2004. Using a difference-in-differences approach, Gershenson found a significant decrease in teacher absences of 0.6 per year, about 10%, in the year following the treatment of “not meeting AYP.” The effect in Title I schools was even more pronounced, with a decrease of 1.6 absences per year. This finding

contributes to the concept that accountability measures, particularly a failure to meet a certain standard, can directly influence teacher behavior.

The literature on school accountability systems and the associated labels suggest effects related to student performance, the perceptions of people within and around schools, and on teacher turnover. The remainder of the literature review explores the more specific case of accountability systems in which schools receive A-F grades on publicly-available report cards.

### **A-F School Grading Systems**

A-F school grading systems serve as an inherently different form of metric when compared to other forms of data sharing about school performance such as proficiency percentages or ambiguous categorical labels. Most adults educated in the United States received academic performance feedback in the form of letter grades, with the labels of A and F evoking responses that connect to a common experience. School letter grades offer simplicity and lower the “cost” of information by summarizing complex measures of performance into an easily-interpreted scale (Chingos, Henderson, & West, 2012). However, the simplicity of the measure may mask important details about educational value beyond composite test scores (Howe & Murray, 2015) and incentivize a narrowing of focus on certain tested grades and subjects within schools (Figlio & Loeb, 2011).

School Performance Grades in North Carolina take a continuous, composite index of school performance indicators including proficiency percentages on standardized tests and graduation rates and convert the numerical value to one of five discrete letter grades from A-F. Prior research suggests such a system produces sharp cutoffs of perceptions and that small numerical differences that convert to different letter grades influence perceptions of quality (Chingos, Henderson, & West, 2012; Olsen, 2013). For example, if a grading system has a

minimum cutoff of 70 for a B, the difference between a score of 69 and 70 may be quite small, while the corresponding letter grades of C and B imply a much larger performance gap. When put in the context of public information about school quality, the labels associated with discrete performance categories can influence perceptions (Jacobsen, Snyder, & Saultz, 2014).

**Experimental impacts of A-F labels.** Experimental evidence demonstrates that people respond more strongly to A-F performance information than when it is presented in other formats. In a 2014 study that experimentally tested reactions to the A-F grade format, Jacobsen, Snyder, and Saultz created a performance data set for three hypothetical schools – one strong, middle, and weak school. The authors developed metrics that represented the same underlying information in four formats: a numeric performance index, a percentage of students reaching a goal (e.g., proficiency), a categorical performance rating, and an A-F letter grade. Participants were assigned one of the conditions and asked to rate the overall performance of the three hypothetical schools on a seven-point scale. Respondents gave higher overall performance ratings to the “strong” school when in the letter grade condition and correspondingly lower scores for the “weak” school than participants in the other conditions. Their results suggest that letter grades elicit a stronger response than similar information provided in alternative formats. These findings could manifest in the North Carolina system through stronger direct responses to the grade from teachers or through stronger responses to the grade from school administrators and local non-educators that impact teachers indirectly.

Few studies to date have explored the impact of A-F school grades on teachers, particularly those designed to infer causality due to the policy. A study by Ladd and Linderholm (2008) attempted to measure teacher perceptions about classroom behavior based on the letter grade assigned to the school. In their experimental study of 96 Florida preservice teachers, the

authors showed an identical 15-minute video of a classroom designed to contain positive, negative, and neutral student behaviors. Ladd and Linderholm assigned each participant to one of three conditions, telling them prior to watching the video that the school had either a letter grade of A, a letter grade of F, or that the classroom was in a “typical” school. Follow-up *t*-tests to an ANOVA revealed significantly more negative responses for the F group than the “typical” group and more negative responses for the F vs. A group. Conversely, the study found a similar effect for positive behaviors with a significant difference between the A and “typical” groups and more positive responses for the A vs. F group. The results demonstrate that the label placed on a school, particularly in the form of A-F grades, can influence educators’ perceptions of otherwise identical learning environments. The experimental evidence also offers insight into how educators directly respond to school labels, particularly on how they view students within those schools.

**Studies of A-F school grade policies.** Because my study seeks to evaluate the impact of the North Carolina policy, the next section devotes special attention to evaluations of A-F policies. Much of the literature exploring impacts of A-F school-level letter grade accountability systems used data from public schools in Florida and New York City. Exogenous policy shocks, such as the introduction of A-F grades, a change to the formula for calculating grades, or the termination of A-F school grade systems served as natural experiments for policy research that employ causal inference methods to evaluate changes. Most studies in this domain have used regression discontinuity approaches (Chiang, 2009; Rockoff & Turner, 2010; Chingos, Henderson, & West, 2012; Winters & Cowen, 2012) to infer causality about the impact of the assignment of letter grades to schools, due to the sharp cutoffs around grade thresholds.

Grading schools on an A-F scale can impact the public and parental perceptions of the quality of schools both directly (Chingos, Henderson, & West, 2012; Charbonneau & Van Ryzin, 2012; Favero & Meier, 2013; Jacobsen, Saultz, & Snyder, 2013) and indirectly through market impacts (Figlio & Kenny, 2009; Figlio & Lucas, 2004). The labels placed on schools can also impact school changes outside the classroom including school and district leadership and the financial and temporal choices they make (Chiang, 2009). Additionally, these labels can influence changes in teacher perceptions (Ladd & Linderholm, 2008; Favero & Meier, 2013) and teacher turnover (Feng, Figlio, Sass, 2010; Dizon-Ross, 2014). These changes can influence student outcomes through the actions of personnel; however, research also offers insight into the impact of labels on subsequent student outcomes (Figlio & Rouse, 2005; Chiang, 2009; Rockoff & Turner, 2010; Winters & Cowen, 2012). Table 2.1 summarizes relevant literature on A-F policies. The following section contains a detailed review of the empirical work exploring these impacts.

Table 2.1  
*Empirical Studies of A-F School Grades Literature Summary*

Topic	Authors	Location	Method	Outcome(s)	Findings
Public perceptions	Chingos, Henderson, & West (2012)	Florida	Regression; RD	Perceptions of school quality	Citizens formed judgments about their public schools in line with letter grades assigned
	Figlio & Kenny (2009)	Florida	RD	Reported donations from community to public schools	Receiving D or F grades negatively impacted school's ability to receive contributions from the community
	Figlio & Lucas (2004)	Florida	RD	Housing values	Significant increase in home values of A vs. C schools in first year of grading policy
Parental perceptions	Charbonneau & Van Ryzin (2012)	New York City	Regression	Parental satisfaction with schools	Parental satisfaction aligned most closely to the raw student performance letter grade
	Favero & Meier (2013)	New York City	Regression	Parental satisfaction with schools	Parental satisfaction aligned with letter grades related to school quality
	Jacobsen, Saultz, & Snyder (2013)	New York City	RD	Parental satisfaction with schools	Lower grades despite little change in actual performance led to declines in satisfaction
Principal practices	Rouse et al. (2013)	Florida	RD	School-level practices	Schools receiving Fs reported lengthening the amount of time for instruction, schedule changes, increased teacher resources, and decreased principal control
	Chiang (2009)	Florida	RD	School spending	Moving a school to "threatened" status with the receipt of an F led to increased spending in curriculum and pedagogical reform

Table 2.1 (continued)

Student achievement	Figlio & Rouse (2005)	Florida	RD	Reading and math achievement	Students made significant gains on state tests the year following an F grade
	Chiang (2009)	Florida	RD	Math achievement	Students made significant improvement in middle school math after attending low-performing elementary schools
	Rockoff & Turner (2010)	New York City	RD	Reading and math achievement	In the period following a low school grade, math and English achievement increased
	Winters & Cowen (2012)	New York City	RD	Reading and math achievement	Students in schools receiving an F had subsequent improvement in English and math in the second year following the F grade
	Winters (2016)	New York City	RD	Reading and math achievement	In the absence of letter grades, subsequent student performance did not improve
Teacher outcomes	Favero and Meier (2013)	New York City	Regression	Teacher satisfaction with schools	Teacher satisfaction had a significant, positive relationship with A-F school grades
	Feng, Figlio, & Sass (2010)	Florida	RD	Teacher turnover	Teachers were more likely to leave schools with a lower grade after changes to the grading formula
	Dizon-Ross (2014)	New York City	RD	Teacher turnover	Teachers were less likely to leave schools receiving an F grade

*Impacts on public perceptions.* Chingos, Henderson, and West (2012) empirically tested the impact of public accountability systems on citizens' perceptions of school quality in Florida. Although the schools received A-F grades beginning in 1999, the study exploited a natural experiment in a 2002 change to Florida's grading formula in which the state added student-level growth to the calculation. This adjustment caused many schools' grades to change unexpectedly from what the grade they would have received under the previous formula. The authors used responses to an oversample of Florida residents to the 2009 Education Policy and Governance Survey in which participants assigned their local elementary, middle, and high school a letter grade of A-F. A regression discontinuity analysis sought to infer causality of the impact on the state-provided A-F grades on the grades respondents assigned to their local schools. The authors found no significant difference at the A/B cutoff but found a marginally significant impact at the B/C cutoff ( $p < .10$ ). Their study concluded that the grades citizens assigned to their local schools reflect the information that is publicly available about education. In other words, people made informed judgments about the quality of their schools using the letter grade in school report cards. In addition, the results suggest an impact on perceptions when schools receive a B grade vs. a C grade, a threshold tested in my study.

Public perceptions and associated behaviors can also be observed indirectly. Two relevant studies from Florida – exploring stated financial support for schools and through market impacts on housing prices – suggest that the grades assigned to schools affected the public. Figlio and Kenny (2009) used the same 2002 change to the Florida grading formula as Chingos, Henderson, and West (2012) to understand the impact of A-F school grades on private contributions to public schools. As data were not directly available on donations and other sources of non-governmental revenue, the authors used a longitudinal survey of Florida

elementary and middle school principals in 2002 and 2004 to examine the impact. Figlio and Kenny (2009) utilized a single survey item attempting to capture contributions: “Approximately how much additional revenue does this school raise annually through other sources of income (e.g., PTAs, community, or business sponsorship, athletic or parking fees, etc.)?” When controlling for year and site fixed effects and various other covariates, the study found significant declines in the reported revenue schools receiving D or F grades in the pre-post period of the grading formula change. Their research suggests that the grade a school receives can influence behavior in addition to perceptions.

The literature also suggests that the letter grades can impact housing markets. Figlio and Lucas (2004) demonstrated that the initial assignment of a school letter grade had implications for Florida housing prices. The period covered by the research coincided with the first three years of Florida’s A+ Schools Initiative by looking at neighborhoods, linking elementary schools and their accountability grades in 1999, 2000, and 2001 to the housing prices. Through the end of 1999, the first year of letter grade assignment to schools, the authors found a 35.6 percent increase in home prices for A schools as compared to C schools, controlling for site and year fixed effects and several other school and real estate covariates. By 2002, however, the impact diminished to a non-statistically significant difference. The authors concluded that despite observing initial impacts of school grades on housing price, the high levels of year-to-year fluctuation in the letter grades did not see the effect persist over time.

The Figlio and Lucas (2004) study may provide evidence to expect an *initial* impact on working conditions that may diminish after the initial shock of the A-F label. This is important because the study of North Carolina looked at the impact of letter grades within the first two years of assignment. Even if the A-F label produced diminished impacts over time, the study

should have detected these initial impacts. Also of note, the current North Carolina accountability design around School Performance Grades is most like the way the Florida system functioned from 1999-2001, when the letter grades were focused mainly on test score levels, not emphasizing progress schools made in student growth.

***Parental satisfaction with public schools.*** In addition to the greater community, public school parents, an important constituency for observing impacts of school performance measures, make judgments about the satisfaction with their school in line with the officially measured performance (Charbonneau & Van Ryzin, 2012). From 2007-2013, New York City determined A-F school grades based on student performance, growth, and measures of school climate. During this period, New York City also annually surveyed teachers, students, and parents. Charbonneau and Van Ryzin (2012) studied the connection between parental satisfaction with their schools from and actual school performance using data from the 2007-08 school year. The authors used a regression analysis to explore the degree to which performance measures predicted parental satisfaction, controlling for school and student characteristics. The outcomes included all summary measures used to compute the A-F grade on the school report card. Of the four measures tested, the student performance score, representing the raw student performance on standardized tests, was the greatest predictor of parental satisfaction with a substantively large effect size of .30 *SD*. The quality review score, derived from survey data and state accountability status, also predicted parental satisfaction with a smaller effect size. The student progress score, determined from the value-added student growth measure, was not a significant predictor of parental satisfaction. The differential associations between measures and parental satisfaction has implications that parents' assessments of quality align better to raw performance than growth. In the case of North Carolina, whose School Performance Grades are based more heavily on student

performance than on growth, there are implications that negative grades could produce a strong negative reaction from parents.

Favero and Meier (2013) also studied New York City survey results from 2007-2009, looking specifically at items related to parent and teacher satisfaction; teacher results will be discussed in a later section. The authors' conceptual framework described that the opinions people form about schools contain both measured and unmeasured school characteristics, an important distinction when studying how school accountability measures influence perceptions. In line with Chingos, Henderson, and West (2012) and Charbonneau and Van Ryzin (2012), the authors concluded that parents' levels of satisfaction with their school aligned closely to the measures of school quality on report cards including student test performance, attendance rates, and school violence.

In 2009, the chancellor of NYC schools raised the requirements for receiving an A grade to continue motivating schools to higher performance (Jacobsen, Saultz, & Snyder, 2013). Thus, a natural experiment occurred when re-norming the grade system, like adjustments made to grade thresholds in Florida in 2002. Following this adjustment, 71% of schools declined by at least one letter grade from the previous year without significant changes in criterion-referenced performance. Jacobsen, Saultz, and Snyder (2013) exploited this natural experiment to understand how the letter grades assigned to a school impacted parental satisfaction. The authors found that the lower reported levels of achievement because of higher expectations led to a decline in parent satisfaction. The introduction of lower grades under an A-F system, despite little to no change in actual academic performance, negatively influenced parental perceptions of their schools in NYC. The results support the possibility that the introduction of A-F grades in

North Carolina could negatively impact perceptions, even with no actual changes in school performance immediately following the release of A-F grades.

***Impacts on school environments.*** A-F grades can also impact changes within schools. Exploiting the change in Florida accountability data in 2002, Rouse et al. (2013) used different items on the survey of principals in the 1999-2000, 2001-2002, and 2003-2004 school years used by Figlio and Kenny (2009) to measure observed shifts in instructional practices in low-performing schools under the high-stakes accountability regime. Principals reported changes including lengthening the amount of time in the school day dedicated to instruction, changes in scheduling, increased resources available to teachers, and decreased principal control. Although the survey only measured principal perceptions, the study showed that schools responded to the A-F labels in tangible ways that could impact teachers.

Sanctions associated with school performance labels can also change financial decisions in school districts. In Florida from 1999-2006, receiving an F grade for one school year opened the possibility for students to be eligible to receive vouchers for private schools if the school received a second F within the next three years (Chiang, 2009). Using a regression discontinuity of schools on the D/F threshold, Chiang found that a school's move to "threatened" status under the Florida accountability system raised spending on categories related to reform in curriculum and pedagogy. The finding indicates a change in subsequent spending behavior related to low performance.

***Impacts on student outcomes.*** School accountability centers on improving student outcomes (Figlio & Loeb, 2011). Labels placed on schools aim to highlight excellence and to hold schools accountable for raising low achievement. Assigning letter grades as a means of school accountability can have positive consequences. Several studies analyzing A-F systems in

Florida and New York City, the systems in which most A-F policy research has occurred, indicated positive changes in student achievement following the assignment of a failing grade, controlling for other explanatory factors (Figlio & Rouse, 2005; Chiang, 2009; Rockoff & Turner, 2010; Winters & Cowen, 2012). These studies also represent rigorous quasi-experimental evidence of A-F label impacts.

Figlio and Rouse (2005) explored Florida 3-5 grade reading and math standardized test score data from 1995-2000 to understand subsequent student achievement after schools received a failing label. This study took advantage of the period immediately preceding and following the assignment of letter grades to schools, which happened for the first time in 1999. The authors found significant gains on state tests in the year following the receipt of an F grade.

Chiang (2009) looked at Florida students in elementary schools under the threat of sanctions due to low performance in the 2002 school year. Leveraging the change in the Florida grading formula that year, he found that students who had attended low-performing elementary schools under threat of state sanctions showed improvement in math achievement through the first 1 to 2 years of middle school. The study provides evidence that high-stakes accountability systems, with school sanctions attached, can drive improved student outcomes over the short-term and medium-term.

Evidence from New York City's implementation of A-F school grading showed that failing letter grades alone, without clear initial sanctions, also had a positive impact on subsequent student performance and math and English. Rockoff and Turner (2010) studied accountability data from New York City elementary and middle schools, exploiting the pre-post period of the advent of the letter grade policy from 2006-2008. The study design closely mirrored that of Figlio & Rouse (2005), applying the methods from Florida to New York City

data on reading and math achievement for students in grades 4-8. Using a regression discontinuity design, they looked at the schools on the margins between letter grades (e.g., D/F, C/D, B/C). They found that in the period following receiving a low school grade, student achievement in math and English increased.

Winters and Cowen (2012) also looked at academic performance in 4-8 grade English and math in New York City public schools in the period following receiving a letter grade. Their study built upon the findings of Rockoff and Turner (2010) by using an additional year of data to test the same hypothesis. This allowed the authors to test both “initial” effects in the first year following the grade and “persistence” effects in the second year. Employing a sharp regression discontinuity approach, the authors used the receipt of the A-F letter grade in 2007 (the first year they were issued) as the treatment, looking at student-level performance in years immediately prior to and following the grade at each cutoff. Only the D/F threshold yielded any significant impacts. The study found a significant positive impact of attending a school that received an F grade on the students’ immediately following year’s English scores, but a non-significant impact for math scores. In the second year following the F grade, however, students saw a significant positive benefit in both English and math.

The endpoint of an A-F accountability system also offers the opportunity to analyze the impact of the policy. Under new mayoral leadership, New York City ended the practice of assigning letter grades to schools in 2014 (Winters, 2016). The end of this practice, combined with the same underlying measurements released on school report cards in subsequent years, led to a natural experiment isolating the impact of the letter grade on student outcomes. The study conducted by Winters (2016) confirmed findings from previous studies that schools receiving D and F grades had improved standardized test performance in subsequent years. By looking at

schools that would have received an F based on their performance in 2014 but did not actually receive an A-F grade due to the end of the policy, Winters found that, without the label, schools did not show improvement relative to schools that would have earned higher grades. This finding demonstrates that public labeling of schools using an A-F system more strongly impacted performance than the same measures without the A-F label.

***Impacts on teachers.*** Although not studied as frequently as student outcomes, a few prior studies have explored the impact of A-F school grades on teachers. Favero and Meier (2013) explored teacher survey responses from 2007-2009 in the period directly before and after the introduction of A-F grading in New York City. The authors created an “overall teacher satisfaction index” by aggregating six items on the survey and using linear regression to identify which variables were associated with their satisfaction. Teacher satisfaction demonstrated a significant positive relationship with the school report card scores for each performance dimension. These results provided a clear parallel to their assessment of the alignment between parental perceptions and measured academic performance. Although the study design cannot claim causality, there existed a correlation between the performance measures on the school report card and teachers’ satisfaction with their schools.

A-F school grades can also impact teacher turnover, but the conclusions are mixed. Feng, Figlio, and Sass (2010) studied the exogenous shock in 2002 in Florida when some schools received higher and others lower performance grades due to a change in the calculation. The study found that teachers were more likely to leave schools whose letter grades dropped because of the updated calculation. Dizon-Ross (2014), however, found that New York City schools with lower accountability grades had *lower* teacher turnover than prior to receiving the low grade. The Dizon-Ross (2014) findings run contrary to Feng, Figlio, and Sass’s study of Florida in 2010.

The effect was more pronounced for teachers deemed high-quality, as their turnover rate was smaller than the general population of teachers in low performing schools.

**Summary of A-F Research.** Although studies of A-F school grading systems have been concentrated in Florida and New York City, the results show some clear patterns that inform the theory of action. A-F grades have consistently shown higher subsequent student achievement, particularly for schools receiving failing grades; this finding was consistent for systems mainly driven by stigma and by those with sanctions that accompanied the failing label. A-F impacts also appear indirectly for non-educators through parental school choice decisions and on local markets through real estate prices. The impact on teachers is not as well known, although the grades a school receives correlate with teacher work environment perceptions and may impact teacher turnover, although the direction of the impact was not clear through conflicting evidence from Florida and New York City.

***Research on A-F Grades in North Carolina.*** Given the nascence of the School Performance Grade policy, little research exists to date evaluating the impacts of the letter grades in North Carolina. Pierson et al. (2015) provided a critique of the policy, citing that the grades in North Carolina did not adequately portray growth or offer parents with actionable school options based on the results. The authors also noted high correlation between student poverty and the grades a school received.

Smith and Imig (2017) also offered several critiques about North Carolina School Performance Grades, using a survey of and interviews with a sample of principals after the receipt of the grades to understand their perceptions. Qualitative responses from principals indicated that the main benefits of A-F accountability grades included the ease of understanding, increased accountability, the possibility of spurring progress, and value to politicians. These

benefits closely matched those addressed by proponents in the literature (Howe & Murray, 2015). The survey also asked principals to indicate the greatest limitations to the A-F grading system. The common themes included inaccuracy/oversimplification, giving a negative view of the school to the community, the grade simply measured poverty levels, a lack of emphasis on growth, and a negative contributor to teacher morale and stress.

The only study to date linking North Carolina School Performance Grades to the Teacher Working Conditions survey came from a 2018 report from the New Teacher Center. The report described a correlation between the letter grades and the various constructs measured by the survey, underscoring the importance of exploring the relationship between teacher working conditions and school accountability. The design of the study, however, conflated the grade and the score together, a flaw addressed by my study.

## **Conclusion**

The advent of the North Carolina School Performance Grades offers an important opportunity to analyze the impacts of a statewide policy with national implications. The assignment of A-F grades to schools in North Carolina represents the latest shift in over 20 years of high-stakes accountability in the state. The letter grade label offers informational accessibility to a larger proportion of the population but runs the risk of oversimplifying and misrepresenting school quality (Howe & Murray, 2015; Tanner, 2016). The research reviewed in Chapter 2 offered evidence that School Performance Grade stands apart as an inherently different type of school accountability measure because the labels are simple, recognizable, and potentially polarizing. Understanding the introduction of A-F school-level performance measurement in North Carolina fills an important gap in the literature.

North Carolina joined 16 states and New York City as large school systems to have instituted A-F grading; the literature, however, offers limited understanding of the impacts of the grades on teachers. The previous research focused on a narrow band of settings, primarily Florida and New York City. Although prolific studies exist on school accountability, the impact of A-F grades remains understudied given the large number of settings in which the grades are used. Thus, the study fills another gap in the literature by exploring A-F grading in a different state.

Finally, prior research on A-F grading focused heavily on public and parental perception and student outcomes. Although some studies offered insights on the impact of A-F school grades on personnel, the literature contains a gap of clearly understanding how the A-F grade influences teacher perceptions. The biennial data from the North Carolina Teacher Working Conditions Survey offer an opportunity to examine changes in teacher responses to relevant items in the setting before and after the state assigned School Performance Grades. The study design described in Chapter 3 aimed to address these gaps by providing evidence of the impact of the A-F School Performance Grade on teachers' perceptions of their schools and subsequent turnover.

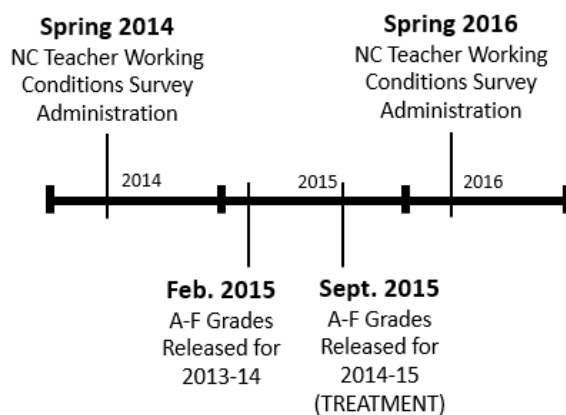
### CHAPTER 3

The study design exploited a natural experiment due to the design of the assignment of A-F School Performance Grades. Schools with similar performance on opposite sides of the grade thresholds created natural treatment and comparison groups to explore the impact of the letter grade label. As discussed in Chapter 2, many studies evaluating similar policies measured student-level outcomes or looked at individual property values at grade cutoffs, naturally lending themselves to a regression discontinuity design (Figlio & Lucas, 2004; Chiang 2009; Rockoff & Turner, 2010; Winters & Cowen, 2012). Likewise, the format of North Carolina School Performance Grades draws distinct cutoffs between continuous composite scores, in which a small difference in numeric value can make a substantial change in the letter grade assigned. The adherence of the letter grades to the performance composite formula made a sharp RD design appropriate for the study.

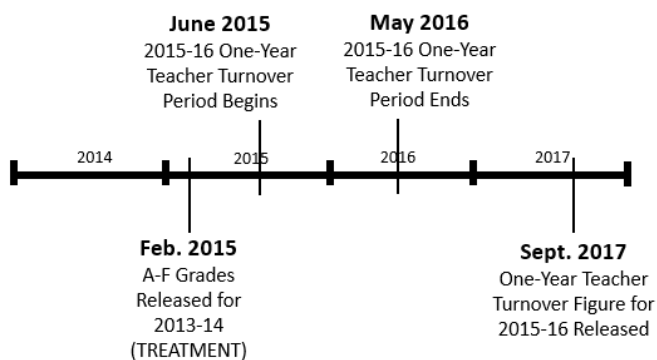
I used the receipt of two different School Performance Grades depending on the research question. For the outcomes coming from NC Teacher Working Conditions 2016 survey data in RQ1 and RQ2, the receipt of a grade for the 2014-15 school year, the most recent grade posted at the time of the 2016 NC Teacher Working Conditions Survey, determined the treatment variable. For RQ3, which explored the impact of the School Performance Grade on actual school-level teacher turnover in 2015-16, I used the grade for the 2013-14 school year to determine the treatment variable (recall that the initial A-F grades for the 2013-14 school year were not released until February 2015).

Teachers responded to administrations of the North Carolina Teacher Working Conditions Survey every two years in the spring from 2002 to 2016. Prior year school-level NC Teacher Working Conditions Survey results from 2014 served as “pretest” covariates. I included

other school-level student and staff characteristics as covariates to increase statistical power. The models also included school-level covariates such as staff turnover, teacher characteristics, and student demographics for the 2013-14 school year to improve statistical power. All these covariates, including the 2014 survey data, were collected prior to the initial release of the A-F letter grades in February 2015. Figures 3.1 and 3.2 show the events that contribute to the data in this study.



*Figure 3.1.* Timeline of events for the study outcomes related to the NC Teacher Working Conditions survey.



*Figure 3.2.* Timeline of events for the study teacher turnover outcome.

The research design addressed three questions. I assessed each research question separately for three groups – elementary, middle, and high schools:

- 1) Does the School Performance Grade impact teachers' perceptions of a) support, b) autonomy, c) accuracy of state assessments, and d) their schools as good places to work and learn?
- 2) Does the School Performance Grade impact teachers' immediate professional plans in terms of a) intending to remain teaching at their same school the following year and b) intending to leave education entirely?
- 3) Does the School Performance Grade impact subsequent teacher turnover?

A sharp regression discontinuity design using data from either side of the grade cutoff line for a “passing” (A, B, or C) vs. “failing” grade (D or F) isolated the effect of the receipt of the School Performance Grade on teacher perceptions of their schools and on teacher turnover. In addition, I explored the impact of three individual grade thresholds – B/C, C/D, and D/F for the elementary and middle school samples. For the high school sample, I explored the impact at the B/C and C/D thresholds only, because there was not a sufficient number of high schools receiving an F to determine an impact at the D/F threshold. For all samples, an insufficient number of schools received A grades to assess the impact at the A/B threshold.

For the first research question, I explored teachers' perceptions of four constructs based on items from the NC Teacher Working Conditions Survey that are conceptually related to the letter grade label given to their schools. For the second research question, I analyzed the teacher-level responses to the NC Teacher Working Conditions Survey item that asked teachers to “describe their immediate professional plans.” For the third research question, I looked at actual

school-level teacher turnover rates in the first year of data available following the receipt of the 2013-14 grade.

### **Analysis**

**Sample selection.** I reduced the total pool of potential schools based on various criteria. I initially considered the full sample of 2,589 schools in the School Performance Grade 2014-15 file. I first eliminated all schools that did not have a valid School Performance Grade for the 2014-15 school year. Removing these 143 schools reduced the sample to 2,446. I then removed all remaining charter schools from the sample. Removing these 142 schools reduced the sample to 2,304.

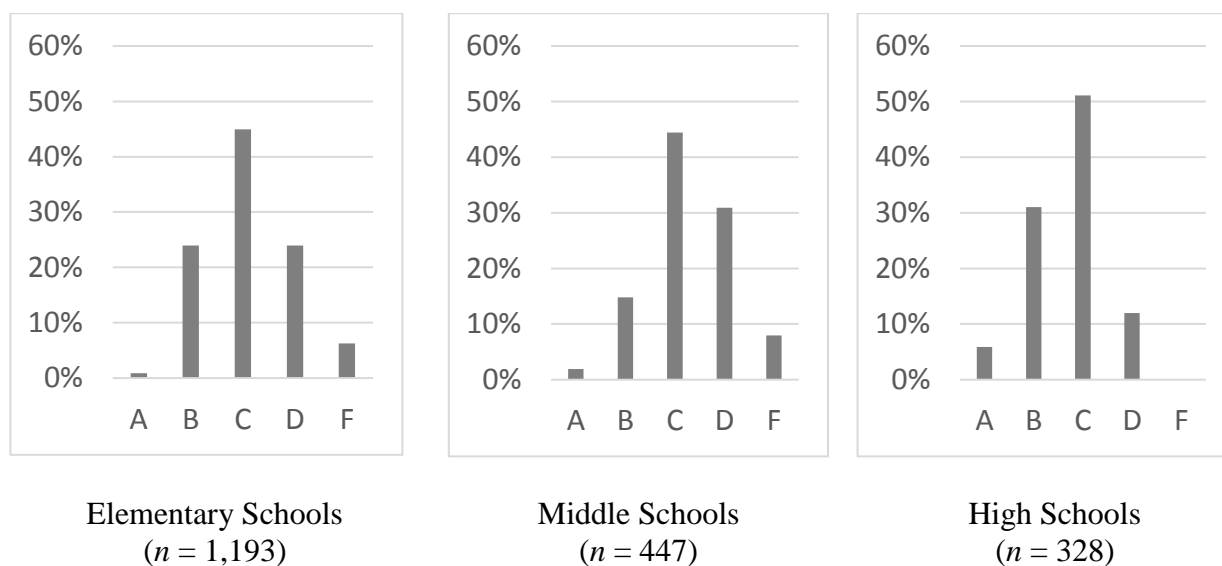
Teacher Working Conditions data are made publicly available online at the school level when at least five teachers responded and the response rate is greater than 40% (New Teacher Center, 2014). Following this guidance, I eliminated schools from the sample that did not have at least five teacher responses. To handle missing data for the teacher turnover outcome and all predictor variables, I used listwise deletion, eliminating schools with missing data for any variable in the model. Removing schools with inadequate Teacher Working Conditions data and missing covariates removed an additional 240 schools from the potential sample, reducing the potential analytic sample to 2,064 schools.

As discussed in Table 1.3, the measures used to determine a School Performance Grade (SPG) depend on the grade levels served by a school. In addition to value-added growth from the EVAAS system, the formula consists of eleven academic performance indicators averaged to determine a letter grade. Due to these differences, I answered each research question separately for three mutually exclusive samples. I based the samples on the measures included in the School Performance Grade score rather than by the grade spans served by the school.

The first sample consisted of schools with a reported performance composite for math, reading, and science end-of-grade testing only. This sample, comprised of elementary and intermediate schools, contained 1,193 schools. For the remainder of the study, I describe this group of schools as the elementary school sample. The second sample contained schools with data for the Math I end-of-course test in addition to scores for math, reading, and science end-of-grade testing, a characteristic associated with middle schools. This second sample accounted for 447 total schools in the middle school sample. The third sample consisted of schools with all seven high school indicators including Math I, English II, and Biology end-of-course tests, Math III participation, ACT scores, WorkKeys test results, and four-year cohort graduation rates. This high school sample, comprised of schools with valid measures for all seven high school outcomes and no elementary or middle school outcomes, contained 328 schools. The three samples included a total of 1,968 schools. This excluded an additional 96 schools from the pool of 2,064 due to a configuration of metrics contributing to the School Performance Grade Score that did not meet the criteria to be included in the elementary, middle, or high school samples.

**Distributions of each sample.** Initial analysis of the 2014-15 School Performance Grades showed different distribution shapes when comparing the three groups of schools, validating the decision to analyze them separately. The elementary school sample demonstrated a slightly skewed left distribution with a higher proportion of schools receiving F grades than A grades. The middle school sample was skewed left with a higher proportion of schools receiving D and F grades than receiving A and B grades. By contrast, the high school sample was skewed right with a higher proportion of schools receiving A and B grades than D and F grades. The inclusion of high school-specific measures with higher means such as graduation rate and math

course rigor inflated the composite scores for high schools. Figure 3.3 summarizes the distribution of schools by sample.



*Figure 3.3.* 2014-15 NC School Performance Grade distribution by sample.

**Model specification.** The approach involved the analysis of cross-sectional data, using pretreatment student and teacher school-level covariates and 2016 NC Teacher Working Conditions Survey outcomes. In this model, the available prior-year Teacher Working Conditions data from 2014 acted as school-level “pretest” covariates to improve statistical power. Because a regression discontinuity uses values around the running variable as means for assignment to treatment or comparison, utilizing panel data was not necessary; a single observation for each school still allowed for causal inference (Lee & Lemieux, 2010). The design of the study is consistent with that described by Schochet (2008) and Dong and Maynard (2013) – one in which the school is the unit of assignment with no random classroom effects. Based on the equation supplied by Dong and Maynard for Model 5.3 (2013, p. 71), the following model guided the analysis for the first two research questions about teacher perceptions and immediate professional plans:

$$\text{Level 1: } Y_{ij} = \beta_{0j} + r_{ij}$$

$$\begin{aligned} \text{Level 2: } \beta_{0j} = & \gamma_{00} + \gamma_{01}(Treatment)_j + \gamma_{02} \left[ \sum_{n=1}^2 (SPG_{Score_j} - SPG_{Cutoff})^n \right] + \\ & \gamma_{03} \left[ \sum_{n=1}^2 (Treatment)_j * (SPG_{Score_j} - SPG_{Cutoff})^n \right] + \gamma_{04}W_j + \mu_{0j} \end{aligned}$$

A two-level model accurately represents the nested structure of the data. The teacher-level data, however, did not contain any identifying information. It was therefore not possible to link teacher responses across years or add any teacher-level covariates to the Level 1 model. This simplified the model to a means-as-outcomes model by eliminating a  $\beta_{1j}$  coefficient in Level 1 and a  $\beta_{1j}$  equation in Level 2.

$Y_{ij}$  represents the outcome of each scale mean for teacher  $i$  at school  $j$ .  $\beta_{0j}$  contains the Level 2 model with all the available school level covariates and  $r_{ij}$  signifies the residual Level 1 error. In the Level 2 model,  $\gamma_{00}$  represents the intercept and  $\gamma_{01}$  serves as the main effect of interest – the impact of the treatment to school  $j$  on teacher perceptions or immediate professional plans. The *Treatment* variable represents receiving a grade at the low end of the threshold in each analysis (i.e., “fail” vs. “pass,” C vs. B, D vs. C, and F vs. D). The  $\gamma_{02}$  coefficient accounts for the contribution of the running variable in the model, the School Performance Grade composite score that determined the letter grade assigned to a school. The difference between the school score and the cutoff value represents the distance of the school from the discontinuity on the x axis.  $SPG\_Score_j$  represents the scale score from 0-100 a school earned for the 2014-15 school year. The  $SPG\_Cutoff$  differed depending on the threshold in question. To account for nonlinear relationships, the model specification includes linear and quadratic terms representing the distance of the running variable from the threshold. I compared results for two different polynomial specifications for each RD analysis – one with only a linear term (the linear model) and one with linear and quadratic running variable terms (the quadratic

model). As shown in Table 1.2, a continuous score of 54.5 (which the state rounds up to a score of 55) determined a “fail” vs. “pass”, the minimum score to receive a C grade. Likewise, the value was 69.5 for the B/C cutoff and 39.5 for the D/F threshold. The  $\gamma_{03}$  coefficient accounts for an interaction term between the treatment variable and the distance of the running variable from the cutoff, effectively allowing the slopes to vary on either side of the cutoff in each model.  $W_j$  and the corresponding  $\gamma_{04}$  coefficient represents school-level covariates. Finally,  $\mu_{0j}$  signifies the school-level mean for each outcome.

For the third research question, I only used school-level data to assess teacher turnover. The following one-level model guided the analysis:

$$Y_j = \beta_0 + \beta_1(Treatment_j) + \beta_2 \left[ \sum_{n=1}^2 (SPG_{Score_j} - SPG_{Cutoff})^n \right] + \beta_3 \left[ \sum_{n=1}^2 (Treatment_j) * (SPG_{Score_j} - SPG_{Cutoff})^n \right] + \beta_4 X_j + r_j$$

For this single-level model,  $Y_j$  represents the school percentage of teacher turnover in the 2016-17 data set for school  $j$ ,  $\beta_0$  represents the intercept,  $\beta_1$  represents the main effect of receiving a grade on the lower side of the threshold of interest,  $\beta_2$  represents the coefficient for the distance of the running variable from the cutoff,  $\beta_3$  accounts for the interaction between the treatment and the distance of the running variable from the cutoff, and  $\beta_4$  and  $X_j$  respectively represent school-level covariates, including school characteristics and teacher turnover in years prior to treatment.

Prior to running analysis, I centered all model covariates to have a mean of zero as part of creating the data set using the *scale* function with the `scale = FALSE` option in the base R package. For the Teacher Working Conditions survey scales in RQ1, I standardized the variables

to have mean zero and standard deviation of one using the *scale* function. Standardizing the outcome variables made it easier to interpret the coefficient values in the model outputs for RQ1. For variables related to immediate professional plans and school-level teacher turnover in RQ2 and RQ3, I did not standardize the outcome variables so that model coefficients could be interpreted as percentage point differences.

***Determining impact for multi-level models.*** As described in the section about the model specification, to appropriately represent the nested structure of the survey data, I created means-as-outcomes models with HLM, using the *lmer* package in R. For each analysis, I included the appropriate scale or item from the Teacher Working Conditions survey as the outcome variable. Each model included multiple specifications of the running variable. In addition, each model contained the relevant “pretest” covariate, defined as the 2014 school-level mean for the corresponding scales or items. Models also included additional school-level covariates from the pretreatment year (2013-14) about student demographics, teacher experience, school size, and principal turnover. Following the HLM estimation of the coefficients, I used the *pt* function in the base R package to determine the *p* value from each *t* statistic to determine statistical significance.

Each model also included weights that align to two different kernels – rectangular and triangular. The rectangular kernel equally weighs all observations regardless of distance from the cutoff value. By contrast, models with a triangular kernel place more weight for observations close to the cutoff than those far away (Lee & Lemieux, 2010). Specifying the different weighting schemes for each analysis strengthened the potential robustness of results, eliminating potential alternative explanations, particularly dealing with results due to random chance. The *lmer* package did not include a built-in mechanism to apply a triangular kernel. To apply the

kernel, I used the *kernelwts* function within the *rdd* R package (Dimmery, 2017) to compute a weight for each observation according to the distance of the running variable from the RD threshold. I then incorporated these weights into the HLM models that used a triangular kernel by specifying the values under the *weights* parameter in each *lmer* model.

***Determining impact of single-level models.*** I assessed the actual teacher turnover outcome in RQ3 as a single-level value model. I used the *lm* command for standard linear regression in the base R package. Following the estimation of the coefficients, I used the *pt* function in the base R package to determine the *p* value from each *t* statistic to determine statistical significance. I used the same *kernelwts* function to determine the regression weights used to compute the models with a triangular kernel.

**Assumptions of RD.** In addition to the assumptions of linear regression, RD models must meet additional criteria, outlined by Lee and Lemieux (2010). First, assignment around the cutoff must be “as good as random.” The design met this assumption because the composite index determines the School Performance Grade; all schools within a certain composite index score range received the same grade based on annual measures that can fluctuate from year to year. In theory, schools could have manipulated different components of the grade, such as inflating a graduation rate, but the aggregate of components from test scores and other measures in the composite makes the possibility of full manipulation unlikely. Next, there must exist no discontinuity at the cutoff using outcomes prior to treatment. I tested this assumption by checking for any discontinuities with covariates as outcomes. None of the pretest covariates for any model or sample demonstrated a significant treatment effect when placed into their corresponding models as outcomes.

The Imbens-Kalyanaraman (2009) optimal bandwidth procedure yielded a bandwidth with too few data points for analysis. Instead of employing this procedure, I used all data points on either side of the pass/fail threshold for each analysis. For the analyses at the B/C, C/D, and D/F thresholds, I set the bandwidth to 10 and 15 points to check for robustness of the impacts to the distance from the grade thresholds.

**Power analysis.** I conducted a power analysis to determine whether a regression discontinuity was feasible for answering the research questions. The parameters used for power analysis required many different calculations for each sample, outcome, and bandwidth. Conducting all power analysis within the PowerUp spreadsheet was challenging to keep track of all the parameters. To address the issue, I wrote functions in R to conduct the RD power analysis for a variety of parameters associated with the different analyses. I built the functions from equations in the PowerUp spreadsheet and from Schochet's 2008 paper on the RD Design Effect. Appendix A includes more detail about the power analysis conducted for the study.

**Model covariates.** Table 3.1 contains the non-standardized descriptive statistics for each sample.

Table 3.1  
*Descriptive Statistics for each Sample*

	Elementary Sample ( <i>n</i> = 1,193 schools)		Middle Sample ( <i>n</i> = 447 schools)		High Sample ( <i>n</i> = 328 schools)	
	Mean	SD	Mean	SD	Mean	SD
SPG Score 2014-15 (Treatment)	60.3	12.4	57.8	12.5	66.4	9.8
Teachers w/ 0-3 Yrs Exp (%)	22.1	11.7	24.0	11.4	22.6	8.6
Teachers w/ 4-10 Yrs Exp (%)	26.6	8.9	24.5	8.3	23.0	6.2
Teachers w/ 11+ Yrs Exp (%)	51.3	13.8	51.6	13.0	54.4	10.1
Number of Teachers	33.5	11.9	41.6	14.8	65.3	28.9
Teachers w/ Adv Degree (%)	28.5	10.8	26.9	10.3	24.3	8.7
Teacher Turnover (%)	13.2	8.2	15.1	8.3	15.0	6.5
American Indian Students (%)	1.8	8.2	1.2	5.4	1.7	7.0
Asian Students (%)	2.1	3.6	2.0	3.2	1.7	2.4
Hispanic Students (%)	16.0	13.1	13.1	10.2	10.3	7.8
Black Students (%)	25.7	23.1	25.4	21.5	27.1	22.6
White Students (%)	50.5	27.8	54.7	25.3	55.8	25.3
Two or More Races Students (%)	3.8	2.1	3.5	1.8	3.2	2.1
Pacific Islander Students (%)	0.1	0.3	0.1	0.2	0.1	0.1
Econ Disadvantaged Students (%)	67.0	23.7	59.4	19.6	51.5	16.9
Support School Mean 2014	3.0	0.3	2.9	0.3	2.9	0.3
Autonomy School Mean 2014	3.1	0.3	3.0	0.3	3.0	0.3
Assessment Accuracy Mean 2014	2.4	0.3	2.3	0.3	2.2	0.2
Good Work/Learn Mean 2014	3.2	0.4	3.1	0.3	3.1	0.3

**Treatment variable.** The release of two School Performance Grades, for 2013-14 and 2014-15, between the 2014 and 2016 NC Teacher Working Conditions Survey administrations presented an analytic challenge. The potential treatments could involve only the initial letter grade assigned in 2013-14, only the most recent letter grade assigned in 2014-15, or a combination of the two letter grades. To better understand the issue, Table 3.2 shows a transition matrix illustrating the percentage of schools moving between grade labels between 2013-14 and 2014-15. For example, in the fourth row and third column of the table, 6% of the total sample of

schools received a D in 2013-14 and improved to a C in 2014-15. As shown in the bolded totals of the diagonals on the table, 72% of schools received the same grade in 2013-14 and 2014-15.

These measures demonstrate that, at least for the first two years of the School Performance Grade policy, the letter grade assigned to a school was relatively stable. Less than one percent of the schools whose letter grades changed between 2013-14 and 2014-15 changed by more than one point on the scale (e.g. B to D or F to C). I propose a follow-up analysis in Chapter 5 that only includes schools whose grades did not change in the first two years and appear on the diagonal of Table 3.2.

Table 3.2  
*School Performance Grade Transition Matrix from 2013-14 to 2014-15*

		2014-15				
		A	B	C	D	F
2013-14	A	<b>2%</b>	1%	0%	0%	0%
	B	1%	<b>18%</b>	5%	0%	0%
	C	0%	5%	<b>34%</b>	5%	0%
	D	0%	0%	6%	<b>15%</b>	2%
	F	0%	0%	0%	2%	<b>3%</b>

Instead of specifying multiple potential treatments using combinations of one or both grades assigned, the main analysis relied on the most recent label corresponding to each outcome. For RQ1 and RQ2, the treatment variable corresponded to the School Performance Grade (SPG) assigned to each school for the 2014-15 school year, the most recent measure prior to the 2016 Teacher Working Conditions Survey administration. The North Carolina Department of Public Instruction released these letter grades in September 2015. For RQ3, the most recent teacher turnover data corresponded to decisions made during Summer 2015. Thus, the 2013-14 School Performance Grade released in February 2015 determined treatment status for RQ3.

For the “fail” vs. “pass” analysis, I generated a treatment value of 1 for all schools that received a D or F and a value of 0 for all schools that received an A, B, or C in 2014-15 for RQ1

and RQ2; I used the same criteria with the 2013-14 grade for RQ3. These data came from the publicly-released data set on the North Carolina Department of Public Instruction website (NC DPI, 2015). In total, 30.2% of the elementary school sample, 38.8% of the middle school sample, and 11.2% of the high school sample received School Performance Grades of D or F and had a treatment value of 1 for the pass/fail analysis. For the additional analysis at the B/C, C/D, and D/F thresholds, schools with a grade on the low side of the cutoff received a 1 and those on the high side received a 0 for the treatment variable.

*Converting the running variable to continuous.* In the administrative data files, the North Carolina Department of Public Instruction rounded each School Performance Grade score to the nearest whole number. Using the integer score as listed in the data file would create challenges with the analysis associated with a discrete running variable. Essentially, each integer value would contain a mass point of results, decreasing statistical power to the point where the analysis could not detect a reasonable effect size. Specifying the running variable as a continuous score assigned unique values to each school, eliminating mass points occupied by multiple schools clustered at the same integer value of the running variable.

I calculated continuous values for the running variable by recalculating the School Performance Grade score from the individual measures. For the performance component of the score, I summed the total number of students passing for each measure divided by the total number of students included for the measure. For the growth component of the score, data obtained from the North Carolina Education Research Data Center (NCERDC) contained a value of the EVAAS growth score rounded to the nearest tenth. I calculated the final continuous running variable by multiplying the performance component score by 0.8 and the growth score by 0.2, the same method used by the state to determine the letter grade assigned to each school. I

then checked to ensure that the continuous value accurately mapped to the integer score in the administrative data set. I used this calculated continuous score in each analytic model instead of the integer value of the running variable.

**Student demographics.** Each model included the percentage of students classified as economically disadvantaged at the school level. NC DPI publishes the total number of students in each school who qualify for free and reduced lunch each year. I included the school-level percentage as a model covariate representing the proportion of economically disadvantaged students.

I also added covariates for the percentage of students representing each racial group as coded by North Carolina. Each model contained the school-level percentage of students from the American Indian, Asian, Hispanic, Black, Two or More Races, and Pacific Islander race groups. I omitted the percentage of White students from the models to avoid multicollinearity. I obtained these percentages from the publicly available Grade, Race, Sex data file, using data from the 2013-14 school year to model pretreatment student demographic characteristics.

**Teacher characteristics.** The models also included demographic data about teachers. NC DPI publishes the percentage of teachers in three experience level bins. Thus, the model included the percentage of teachers in their first 0-3 years of teaching and the percentage of teachers with 4-10 years of experience. Additionally, I added the percentage of teachers with advanced degrees to enhance power related to varying levels of teacher credentials in schools. I also included the number of teachers in each school in the model to represent differences in school size. These data came from the publicly available Personnel data file on the NC DPI website (NC DPI, 2016). Like the student demographic variables, I used data from the pretreatment school year of 2013-14.

**Staff turnover.** Each model also contained the percentage of two-year average of teacher turnover at the school level since the last administration of the Teacher Working Conditions Survey. Because of the lag in the teacher turnover data, the appropriate values for the two-year average correspond to the 2013-14 and 2014-15 school years from the publicly available Personnel data file.

Principal turnover also has the potential to greatly impact teacher perceptions of their schools (Burkhauser, 2017). Thus, I included two dummy variables indicating if the school changed principals in either of the two school years between the 2014 and 2016 Teacher Working Conditions survey administrations. Data on principal turnover are not available in the public files on the DPI website. I created variables for principal turnover using the Teacher Pay file from the NCERDC to see whether the ID for the principal changed between 2014-15 and 2015-16 and between 2015-16 and 2016-17. If the personnel ID for the principal did not match for two adjacent years, I assigned a value of 1 for the principal turnover covariate and 0 otherwise.

**Prior year Teacher Working Conditions survey results.** Prior-year results for the outcome variables contributed most to increased power among the model covariates. The lack of identifiable teacher responses did not allow for Level 1 control of previous responses, so I calculated school-level means for each outcome variable from the 2014 NC Teacher Working Conditions Survey administrations for the *Support, Autonomy, Overall Work Climate* and *Immediate Professional Plans* scales and items. Each outcome model contained its corresponding “pretest” covariate from the 2014 survey. Table 3.1 displays the school-level means for these variables.

## **Constructing Outcome Variables**

Outcome variables for the first two research questions came from the NC Teacher Working Conditions Survey results. The state has administered the survey to all teachers in North Carolina every two years since 2002 (Maddock, 2009). Since 2008, the response rate to the survey has been greater than 85%, with over 86% of school-based licensed educators responding to the 2016 survey (New Teacher Center, 2018). Thus, non-response bias most likely does not pose a significant threat to inference (Groves et al., 2009).

Although the New Teacher Center maintains school-level data for each survey item via its website, I obtained teacher-level responses from the North Carolina Education Research Data Center. The teacher-level data have no personal identifiers attached to preserve anonymity of responses. Although not linkable across years or to teacher-level covariates, the teacher-level data allowed for accurate calculation of reliability and the confirmation of a one-factor solution for each of the items I analyzed. The teacher-level data also accurately reflect the two-level structure of teacher responses to a school-level treatment.

The NC Teacher Working Conditions Survey offered respondents a “don’t know” option for all outcome items for the teacher perceptions constructs. I recoded all “don’t know” responses as missing data before calculating the mean values by scale and school. Thus, I deleted any teacher-level observations from the sample with one or more “don’t know” responses to any item used to determine the outcome variables. Table 3.3 contains the number of missing and don’t know responses for each outcome and sample.

Table 3.3  
*Number and Percentage of Missing and “Don’t Know” Responses*

	Elementary Schools n = 37,861 teachers	Middle Schools n = 16,686 teachers	High Schools n = 18,636 teachers
Support	2,857 (8%)	1,682 (10%)	1,894 (10%)
Autonomy	1,471 (4%)	811 (5%)	891 (5%)
Accuracy of Assessments	2,794 (7%)	1,331 (8%)	1,596 (9%)
Good Place to Work/Learn	846 (2%)	484 (3%)	382 (2%)
Immediate Professional Plans	460 (1%)	227 (1%)	157 (1%)

**Perceived support.** I combined three items from the North Carolina Teacher Working Conditions Survey that correspond to teachers’ perceptions of support. Table 3.4 contains the items that conceptually capture the construct of support, each of which comes from the subscale of Community Support and Involvement on the Teacher Working Conditions Survey. The survey questions each follow a four-point scale from “strongly disagree” to “strongly agree.” I combined the three items into a new scale entitled *Support* by obtaining the mean of numerical codes corresponding to each teacher’s response, assigning 1 = Strongly Disagree, 2 = Disagree, 3 = Agree, and 4 = Strongly Agree. I calculated the mean for the three items for each teacher. The resulting mean score at the teacher level across the three items served as the dependent variable for the analysis of *Support*.

Table 3.4  
*Survey Items for Teacher Perception Outcomes by Construct*

Construct	2016 Scale Reliability	Items*
Support	.87	Parents/guardians support teachers, contributing to their success with students. Community members support teachers, contributing to their success with students. The community we serve is supportive of this school.
Autonomy	.93	Teachers are recognized as educational experts. Teachers are trusted to make sound professional decisions about instruction. Teachers are relied upon to make decisions about educational issues.
Accuracy of State Assessments	--	State assessments accurately gauge students' understanding of standards.
Good Place to Work and Learn	--	Overall, my school is a good place to work and learn.

\*The response scale is a four-point Likert scale with 1 = Strongly Disagree, 2 = Disagree, 3 = Agree, and 4 = Strongly Agree

**Perceived autonomy.** The study also measured the impact of the School Performance Grade on teachers' perceived autonomy. The selected items come from the Teacher Leadership subscale of the Teacher Working Conditions Survey. Collectively, these items relate to teachers' perceptions of autonomy and trust in them to influence their school environment. Like the first research question, the survey items each follow a four-point scale from "strongly disagree" to "strongly agree." I combined the three items into a new scale entitled *Autonomy* by obtaining the mean of numerical codes corresponding to each teacher's response, assigning 1 = Strongly Disagree, 2 = Disagree, 3 = Agree, and 4 = Strongly Agree. I calculated the mean for the three items for each teacher and the mean score for the three items for each teacher served as the dependent variable for the analysis of *Autonomy*.

**Perceived accuracy of state assessments.** The study also assessed the impact of a failing School Performance Grade on teachers' belief about the accuracy of state assessments to

measure students' understanding of the standards. Unlike the previous two outcomes that used scales of multiple items, I answered this question given teachers' responses to the single item, "State assessments accurately gauge students' understanding of standards." The survey item followed a four-point scale from "strongly disagree" to "strongly agree." I looked at the numerical codes corresponding to each teacher's response, assigning 1 = Strongly Disagree, 2 = Disagree, 3 = Agree, and 4 = Strongly Agree. The teacher-level value for this item was the dependent variable for the analysis of the *Accuracy of State Assessments* construct.

**Good place to work and learn.** Receiving a failing School Performance Grade also potentially impacted teachers' overall perceptions of their school as good places to work and learn. I answered this question using teachers' responses to the single item, "Overall, my school is a good place to work and learn." I assigned numerical codes corresponding to each teacher's response, with 1 = Strongly Disagree, 2 = Disagree, 3 = Agree, and 4 = Strongly Agree. The teacher-level response to this item served as the dependent variable for the analysis of the *Good Place to Work and Learn* construct.

**Scale validation.** Creating outcome variables involved combining items into new scales. For *Support* and *Autonomy*, I combined three survey items respectively for the outcome variables. I used the alpha function in the R *psych* package (Revelle, 2017) to determine the reliabilities for the 2016 survey with the sample of schools with complete data across all covariates. As shown in Table 3.4, with values of .87 for the *Support* scale and .93 for the *Autonomy* scale, these scales demonstrated strong reliability. These measures showed stability over time, with the 2014 results demonstrating the same reliability statistics.

Prior to finalizing the outcome variables, I ran an exploratory factor analysis with the chosen items from the 2016 NC Teacher Working Conditions Survey to ensure a factor structure

consistent with the constructs. I used the *factanal* function in R with a varimax rotation and a four-factor solution. Table 3.5 summarizes the factor loadings for all values greater than 0.3 and displays a factor structure consistent with the four scales and items explored in the first research question. In addition, I ran a confirmatory factor analysis with the same structure to ensure goodness of fit using the *cfa* function in the *lavaan* package in R (Rosseel et al., 2017). The Comparative Fit Index of .99 indicated good model fit, as the value was higher than the threshold value of .95 recommended by Hu and Bentler (1999).

Table 3.5  
*Factor Loadings from Exploratory Factor Analysis on 2016 NC Teacher Working Conditions Survey Items*

	Factor 1	Factor 2	Factor 3	Factor 4
Parents/guardians support teachers, contributing to their success with students.	0.67			
Community members support teachers, contributing to their success with students.	0.86			
The community we serve is supportive of this school.	0.85			
Teachers are recognized as educational experts.		0.80		
Teachers are trusted to make sound professional decisions about instruction.		0.91		
Teachers are relied upon to make decisions about educational issues.		0.84		
State assessments accurately gauge students' understanding of standards.			0.43	
Overall, my school is a good place to work and learn.				0.94

\*All factor loadings less than 0.3 were omitted from the table.

**Teacher turnover.** The final two research questions focused on the impact of a failing School Performance Grade on immediate professional plans and actual teacher turnover. I measured this outcome using two data sources. First, teachers indicated their intentions for their immediate professional plans on an item of the Teacher Working Conditions Survey. The item includes six categorical responses, summarized with the state-level 2016 distribution of responses in Table 3.6.

Table 3.6  
*Distribution of Teachers' Immediate Professional Plans, 2016 NC Teacher Working Conditions Survey*

Continue teaching at current school	81%
Continue teaching in district, but leave school	5%
Continue teaching in state, but leave district	3%
Continue working in education, but pursue an administrative position	3%
Continue working in education, but pursue a non-administrative position	3%
Leave education entirely	5%

*Source: NC Teacher Working Conditions Results, 2016*

The second research question assessed planned retention in two areas. For the percentage of teachers indicating they planned to continue teaching at their current school, I coded the teacher-level responses as 0 for indicating that teachers will “continue teaching at current school” and 1 for all other non-missing responses indicating a teacher intends to no longer teach in the school. For the outcome of teachers planning to leave education entirely, I coded the teacher-level responses as 1 for all teachers intending to leave education entirely and 0 for all other responses.

The third research question explored actual teacher turnover rates for the year immediately following the treatment of a failing letter grade from teacher turnover data. These data came from publicly-available school-level means of teacher turnover in the personnel data file published by NC DPI. As referenced in the section about models, the single-level model used the school-level prior-year teacher turnover percentage reported in the 2017 personnel data file, the most recent available at the time of this study.

### **Robustness Checks**

I ran an RD estimator of the impact at each discontinuity using the 2014 NC Teacher Working Conditions Survey pretest covariates as outcomes. The same treatment determined by the 2014-15 School Performance Grade will apply to the schools for this analysis. To validate the

use of the RD to assess impacts in the period after treatment, discontinuities should not exist in the hypothesized direction for each outcome in the period prior to treatment (Lee & Lemieux, 2010). I did not find any significant discontinuities prior to treatment for any of the most recent pretest covariates. This made sense because prior to the release of School Performance Grades, the cutoffs in performance composites had no inherent meaning.

**Multiple comparisons.** To account for multiple comparisons, I applied the standards from version 3.0 of the What Works Clearinghouse (WWC) handbook (2013). Group comparison studies in the WWC require adjustments for multiple comparisons using the Benjamini-Hochberg method with an error rate of 0.05. I adjusted for multiple comparisons for each of the three samples and seven outcomes. For all analyses with significant findings for three or more models, I conducted a multiple comparisons test and included a table with the results in Chapter 4. I sorted each list of  $p$  values from smallest to largest and applied the formula  $\frac{l}{m} * .05$  to calculate the critical value to determine significance, where  $l$  is the rank of the  $p$  value in the sorted list and  $m$  is the total number of comparisons. If any  $p$  value across the study was less than its corresponding critical value, all findings with  $p$  values less than the value meeting the critical threshold would be deemed significant.

**Sensitivity analysis.** The main analysis explored robustness to a few model specifications including two different polynomial specifications, two bandwidths, and two RD kernels for each threshold tested. For findings that were robust to these differences, I ran additional models to further rule out alternative explanations for the findings. I checked three additional bandwidths, within five, seven, and 12 points of the grade cutoffs, to see if the results remained significant.

I executed the analysis aligned to the methods described in Chapter 3. Chapter 4 describes the results from each analysis.

## CHAPTER 4

The following chapter contains the results from the analysis of the impact of the School Performance Grade label on teachers' perceptions and decision to remain teaching in their schools. The main analysis assessed the impact of a pass vs. fail for all three samples. In addition, I conducted assessments of the B/C, C/D, and D/F cutoffs for the elementary and middle school samples and the B/C and C/D cutoffs for the high school sample.

Each section includes scatter plots of the outcome of interest and its relationship to the discrete values of the School Performance Grade score. Following the plots, I present separate tables of results for each outcome and sample that include impact estimates using different specifications of the model functional form, different smoothing kernels around the cutoff points, and different bandwidths. These checks helped to understand the robustness of results. Each section also contains a description of the results and the conclusions drawn for each research question and sample. At the end of the chapter, I present follow-up sensitivity analyses for outcomes that yielded significant results to rule out alternative explanations.

I began each analysis by plotting the mean values of each outcome variable clustered at discrete values of the running variable. Following the advice of Lee and Lemieux (2010), plotting clustered outcome data in bins allows for easier visual inspection of a discontinuity than a scatter plot of outcome and running variable values for each school. Separating the running variable into bins of whole numbers aligns to the format in which North Carolina releases the School Performance Grade scores – nearest whole number values. As defined in Chapters 1 and 3, I chose the threshold between grades of C and D as the cutoff for this analysis. This cutoff corresponds to a School Performance Grade score of 54.5 or higher (55 for whole number values) for a passing grade. Dashed vertical lines in each plot represent the cutoffs between

grades of D/F at 39.5, C/D at 54.5, B/C at 69.5, and A/B at 84.5. Because no schools in the high school sample received an F grade in 2014-15, I omitted the D/F cutoff from all plots of the high school sample.

An initial figure in each section contains the plot of discrete points for all three samples with the results of two local regressions: one for schools receiving a “failing grade” of D or F on the left-hand side of the cutoff and one for schools receiving a “passing” grade of A, B, or C on the right-hand side of the cutoff. I repeated these local regressions for each of the letter grades in separate plots for each sample to visually assess discontinuities at the B/C, C/D, and D/F thresholds. The regression calculates the predicted values of the outcome variable using a linear and quadratic specification of the running variable described by the following equation:

$$Outcome_i = \beta_0 + \beta_1(SPG\_Score) + \beta_2(SPG\_Score)^2$$

A visual vertical gap in the values of the left-hand and right-hand regressions at the cutoff would signal a potential treatment effect to confirm with the full models. In total, I completed twenty-one analyses for pass/fail corresponding to seven outcome variables and three samples. It is important to note that the local regressions contain no additional covariates and thus do not represent the complexity of the full analytic models. The addition of covariates, particularly pretreatment year outcome measures, add additional power to the estimates and may cause the model coefficients to differ in magnitude from the visual inspection.

Following plots for the full sample and local regressions for each individual grade received, the impact estimates are summarized in tables. Each result table summarizes the value of the coefficient and standard error of treatment variable, a binary value indicating that a school received a failing grade (D or F). In addition, each table also contains the teacher-level  $n$  size and school-level  $n$  size (in brackets) for each sample, threshold, and bandwidth. As noted in Chapter

3, I centered all covariates at zero and centered the running variable at the relevant threshold value. In addition, I standardized the mean scores for all items related to teacher perceptions. Thus, for RQ1, the coefficients can be interpreted as standard deviation differences in the outcomes. For RQ2 and RQ3, I left the outcome variables related to immediate professional plans and actual teacher turnover in their original form. This means that the coefficients for these items can be interpreted as proportions of teachers.

For each threshold and bandwidth, I ran the analysis with four different model specifications to test for robustness of results. The first coefficient column represents a linear specification of the running variable; this column is labeled as “linear” to signify the highest-order polynomial term. The second column represents the same analysis with two terms with the running variable for non-linearities – linear and quadratic – labeled as “quadratic” as the highest-order polynomial term. Each of these columns represents models with a rectangular kernel around the cutoff. The third and fourth coefficient columns have the same linear and quadratic maximum terms, but with a triangular kernel at the cutoff.

### **Research Question 1: Teacher Perceptions of Working Conditions in their Schools**

The first research question explored the impact of the School Performance Grade label on teacher perceptions of four aspects of their working environments: a) support, b) autonomy, c) accuracy of state assessments, and d) the extent to which their schools are good places to work and learn. I present the results for each of the four constructs in RQ1 for each of the three samples – elementary, middle, and high schools – in the following section.

**Teacher perceptions of support.** As shown in Figure 4.1, each of the school samples showed a strong, positive relationship between a school’s continuous School Performance Grade score and the school-level mean for perceived support. Each school sample also demonstrated

higher variance between schools at the highest and lowest grades at the end of each distribution than near the center. This heteroscedasticity is not surprising given that fewer schools are clustered at the tails of the distribution. Visual inspection at the threshold between the schools on the passing and failing side of the School Performance Grade for each of the three schools samples showed no distinct evidence of discontinuities.

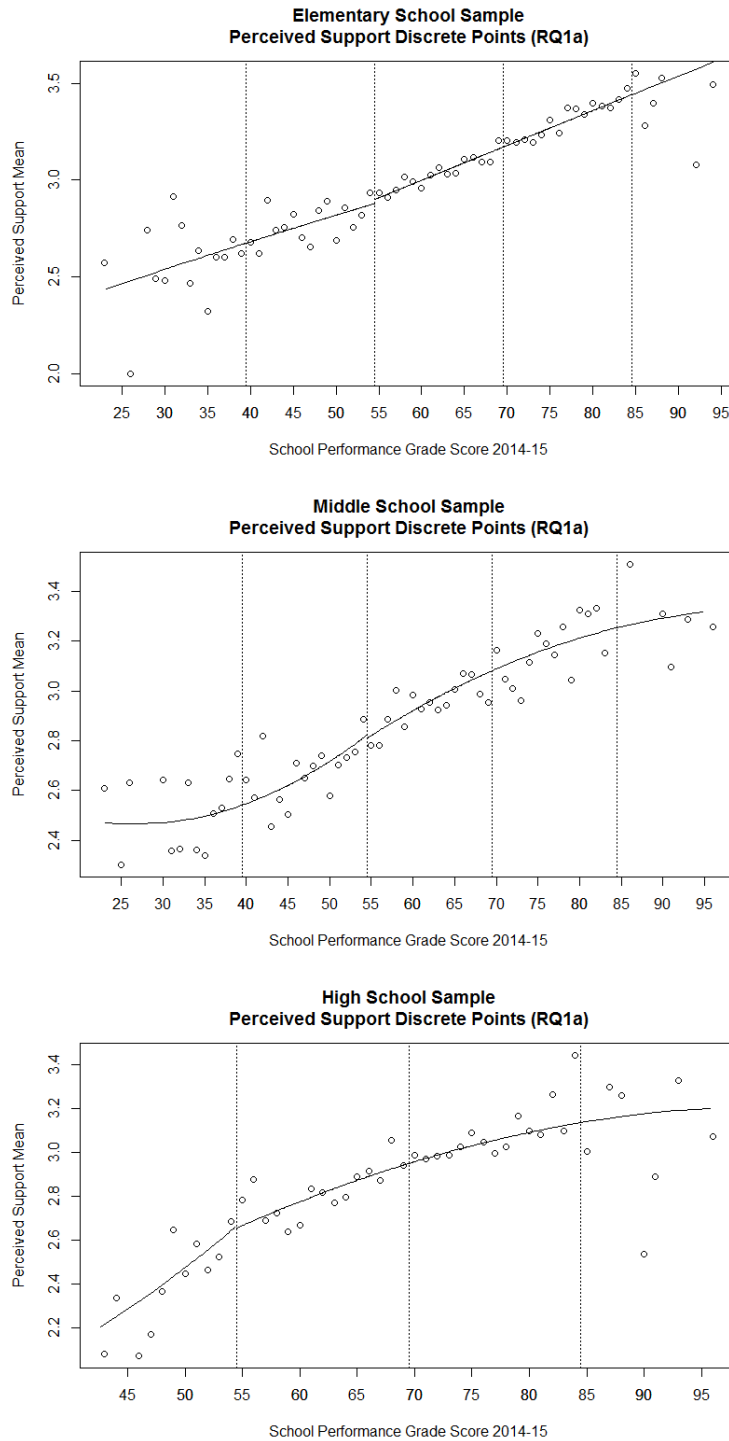
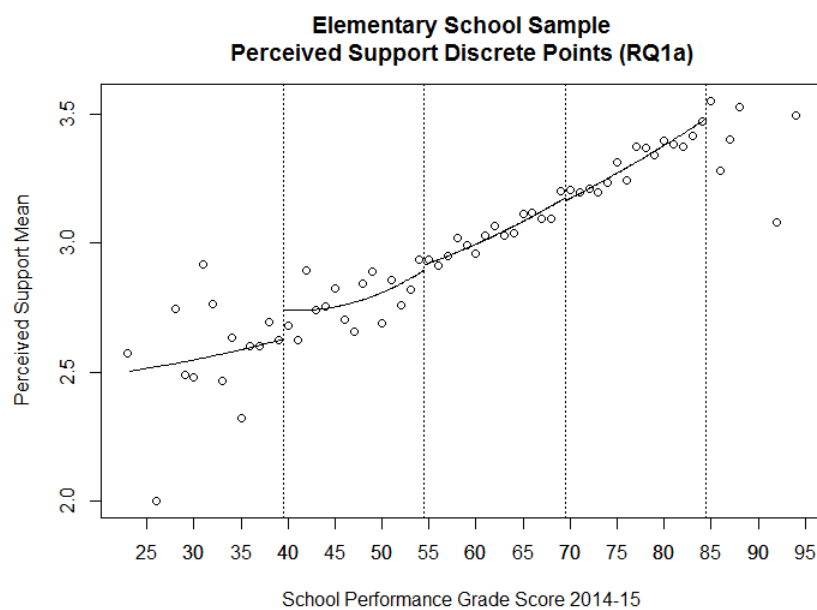


Figure 4.1. Discrete point plots and local pass/fail regressions for perceived support.

**Elementary school sample.** The correlation of values of perceived support and the School Performance Grade score was much greater for schools receiving a grade of B or C than

for schools receiving a grade of D or F. Shown in Figure 4.2, the plot of local regressions for individual grade labels revealed a potential discontinuity at the D/F cutoff.



*Figure 4.2.* Local regressions by letter grade for elementary school perceived support.

Table 4.1 includes the coefficient and standard error values for each model specification, bandwidth, and grade threshold for perceptions of support for the elementary sample. For all models run at individual grade cutoffs, no coefficients were statistically significant. Based on this evidence, it appears that the School Performance Grade label had no significant impact on elementary school teachers' perceptions of support.

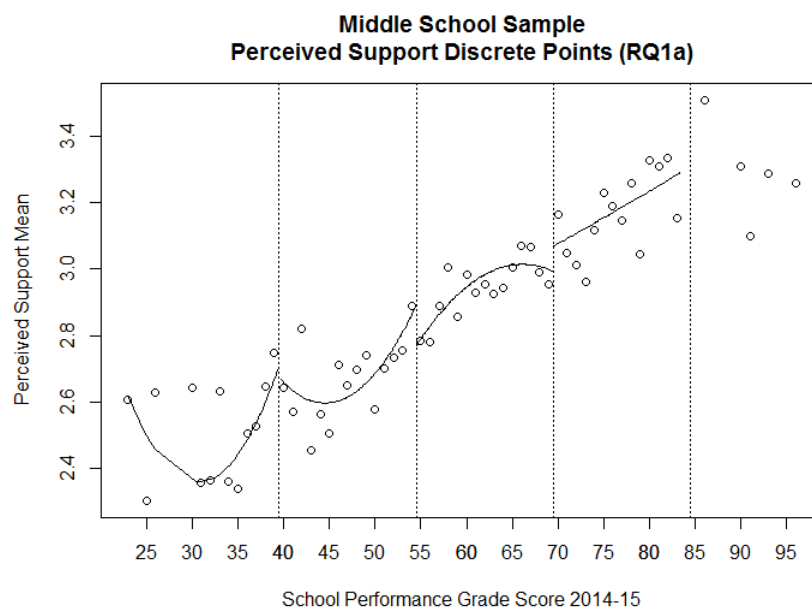
Table 4.1  
*Perceptions of Support – Elementary School Sample*

Threshold	Bandwidth	Sample Size	Rectangular Kernel		Triangular Kernel	
			Linear	Quadratic	Linear	Quadratic
Pass/Fail	All Values	35,004	-0.05	-0.07	-0.05	-0.07
		[1,193]	(0.04)	(0.05)	(0.03)	(0.05)
B/C	10 pt	16,522	0.02	0.01	0.02	0.01
		[562]	(0.05)	(0.07)	(0.05)	(0.07)
	15 pt	24,153	0.02	0.01	0.02	0.02
		[805]	(0.04)	(0.06)	(0.04)	(0.06)
C/D	10 pt	16,425	-0.06	0.08	-0.04	0.09
		[567]	(0.06)	(0.08)	(0.05)	(0.08)
	15 pt	23,485	-0.08	-0.00	-0.07	0.02
		[816]	(0.05)	(0.07)	(0.05)	(0.07)
D/F	10 pt	6,126	-0.12	0.03	-0.08	-0.00
		[226]	(0.11)	(0.15)	(0.11)	(0.15)
	15 pt	9,943	-0.04	-0.16	-0.07	-0.10
		[359]	(0.09)	(0.12)	(0.09)	(0.12)

\* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$ ; Critical values unadjusted for multiple comparisons.

For each analysis, the teacher-level sample size is followed by the school-level sample size in brackets. Each table entry represents the impact estimate of the discontinuity with the standard error of the estimate in parentheses. The entries represent variations in the smoothing kernel and polynomial specification of the running variable, letter grade thresholds, and RD bandwidths.

***Middle school sample.*** Like the elementary school sample, the middle school plot of discrete points demonstrated a strong, positive relationship between a school's continuous School Performance Grade score and the school-level mean for perceived support. The local regression lines for schools receiving a C grade showed a possible discontinuity at the border with both a D and B, but this was likely driven by the local regression curves showing strong quadratic terms for schools receiving an F, D, and C.



*Figure 4.3.* Local regressions by letter grade for middle school perceived support.

In contrast to the visual plot, the regression analysis yielded no significant discontinuities at any grade threshold. Table 4.2 includes the coefficient and standard error values for each model specification, bandwidth, and grade threshold for perceptions of support for the middle school sample. For all models run, none of the results were statistically significant. Thus, for the middle school sample, there was no evidence of an impact of the School Performance Grade label on teacher perceptions of support.

Table 4.2  
*Perceptions of Support – Middle School Sample*

Threshold	Bandwidth	Sample Size	Rectangular Kernel		Triangular Kernel	
			Linear	Quadratic	Linear	Quadratic
Pass/Fail	All Values	15,004	-0.07	-0.02	-0.06	-0.02
		[443]	(0.05)	(0.07)	(0.05)	(0.07)
B/C	10 pt	6,280	0.14	-0.05	0.11	-0.10
		[177]	(0.09)	(0.14)	(0.09)	(0.14)
	15 pt	9,048	0.09	0.08	0.10	0.06
		[256]	(0.07)	(0.11)	(0.07)	(0.11)
C/D	10 pt	7,875	-0.00	0.09	0.02	0.15
		[235]	(0.08)	(0.12)	(0.08)	(0.11)
	15 pt	11,226	-0.02	0.06	-0.01	0.06
		[336]	(0.07)	(0.09)	(0.06)	(0.09)
D/F	10 pt	3,252	0.14	0.08	0.13	0.05
		[105]	(0.16)	(0.24)	(0.15)	(0.23)
	15 pt	5,344	0.08	0.09	0.08	0.15
		[169]	(0.13)	(0.18)	(0.13)	(0.18)

\* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$ ; Critical values unadjusted for multiple comparisons.

For each analysis, the teacher-level sample size is followed by the school-level sample size in brackets. Each table entry represents the impact estimate of the discontinuity with the standard error of the estimate in parentheses. The entries represent variations in the smoothing kernel and polynomial specification of the running variable, letter grade thresholds, and RD bandwidths.

**High school sample.** Like the elementary and middle school samples, the high school plot of discrete points showed a strong, positive relationship between a school's continuous School Performance Grade score and the school-level mean for perceived support. Shown in Figure 4.4, the local regressions suggested a potential discontinuity at the C/D cutoff, particularly when accounting for the quadratic regression term.

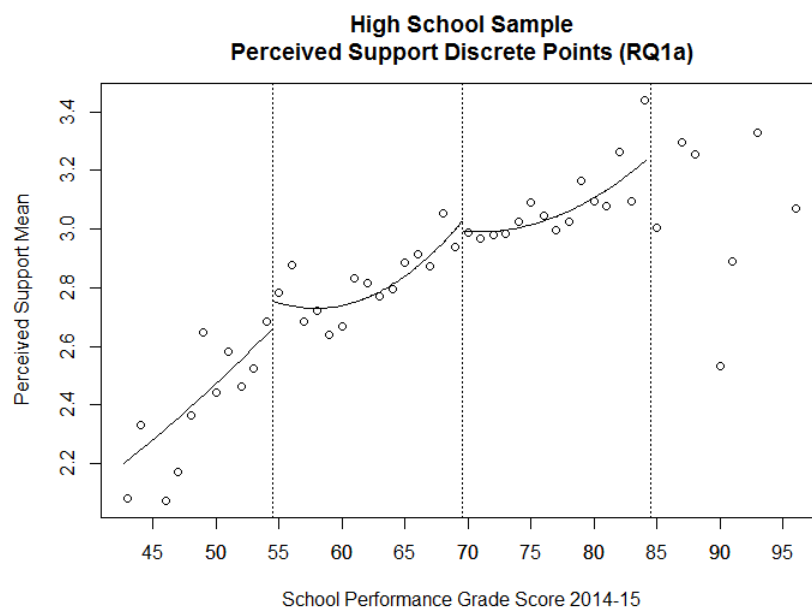


Figure 4.4. Local regressions by letter grade for high school perceived support.

Table 4.3 includes the coefficient and standard error values for each model specification, bandwidth, and grade threshold for perceptions of support for the high school sample. The pass/fail analysis and the assessment of the B/C threshold did not show any significant results. However, all eight tested specifications at the C/D cutoff were statistically significant.

Table 4.3

*Perceptions of Support – High School Sample*

Threshold	Bandwidth	Sample Size	Rectangular Kernel		Triangular Kernel	
			Linear	Quadratic	Linear	Quadratic
Pass/Fail	All Values	16,742	-0.12	-0.18	-0.12	-0.18
		[327]	(0.07)	(0.10)	(0.07)	(0.10)
B/C	10 pt	11,243	0.02	-0.06	0.01	-0.06
		[213]	(0.06)	(0.08)	(0.06)	(0.08)
	15 pt	14,492	-0.03	0.00	-0.02	-0.00
		[274]	(0.05)	(0.07)	(0.05)	(0.07)
C/D	10 pt	6,446	<b>-0.20*</b>	<b>-0.41**</b>	<b>-0.24*</b>	<b>-0.39**</b>
		[137]	<b>(0.09)</b>	<b>(0.12)</b>	<b>(0.09)</b>	<b>(0.12)</b>
	15 pt	10,026	<b>-0.15*</b>	<b>-0.25*</b>	<b>-0.17*</b>	<b>-0.28*</b>
		[205]	<b>(0.08)</b>	<b>(0.11)</b>	<b>(0.08)</b>	<b>(0.10)</b>

\* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$ ; Critical values unadjusted for multiple comparisons.

For each analysis, the teacher-level sample size is followed by the school-level sample size in brackets. Each table entry represents the impact estimate of the discontinuity with the standard error of the estimate in parentheses. The entries represent variations in the smoothing kernel and polynomial specification of the running variable, letter grade thresholds, and RD bandwidths.

Table 4.4 shows the  $p$  values for each of the significant results with application of the Benjamini-Hochberg adjustment. When adjusting for multiple comparisons, these models demonstrated evidence of an impact of the School Performance Grade label on high school teachers at the C/D threshold for four of the eight specifications. These values, ranging from -0.24 to -0.41  $SD$  are larger than the MDES values to their corresponding samples and bandwidths. The sensitivity analysis later in the chapter provides some additional evidence that the findings are consistent with other bandwidths and specifications as well. Although not significant for elementary and middle school teachers, or at the B/C or Pass/Fail thresholds, the analysis suggests that receiving a grade of D versus a grade of C negatively impacted high school teacher perceptions of support.

Table 4.4  
*p* Values and Multiple Comparison Adjustment for High School Perceived Support  
*m* = 20 comparisons

Threshold	Bandwidth	Specification	<i>p</i> Value	Rank	Critical <i>p</i> Value After Adjustment	Significant After Adjustment?
C/D	10 pt	Quadratic, Rectangular	.001	1	.003	Yes
C/D	10 pt	Quadratic, Triangular	.002	2	.005	Yes
C/D	15 pt	Quadratic, Triangular	.008	3	.008	Yes
C/D	10 pt	Linear, Triangular	.009	4	.010	Yes
C/D	15 pt	Quadratic, Rectangular	.022	5	.013	No
C/D	15 pt	Linear, Triangular	.028	6	.015	No
C/D	10 pt	Linear, Rectangular	.032	7	.018	No
C/D	15 pt	Linear, Rectangular	.044	8	.020	No

The Benjamini-Hochberg method involved sorting the *p* values from least to greatest, and assigning a ranking, with 1 representing the smallest *p* value. The rank, number of comparisons, and critical value (.05) are used to calculate an adjusted critical *p* value for each model that determines whether a model is still significant after multiple comparisons adjustment.

**Perceived autonomy.** Each of the samples indicated a weak positive relationship between a school's continuous School Performance Grade score and the school-level mean for perceived autonomy, especially compared to the stronger correlation of the grade with support. Shown in Figure 4.5, visual inspection at the threshold between the schools on the passing and failing side of the School Performance Grade for the elementary and high school samples showed no evidence of a discontinuity. The middle school sample, however, demonstrated a potential discontinuity in the *opposite* direction I expected. In other words, the visual inspection showed potentially that middle school teachers in schools receiving failing grades reported *higher* levels of autonomy than those in schools receiving passing grades.

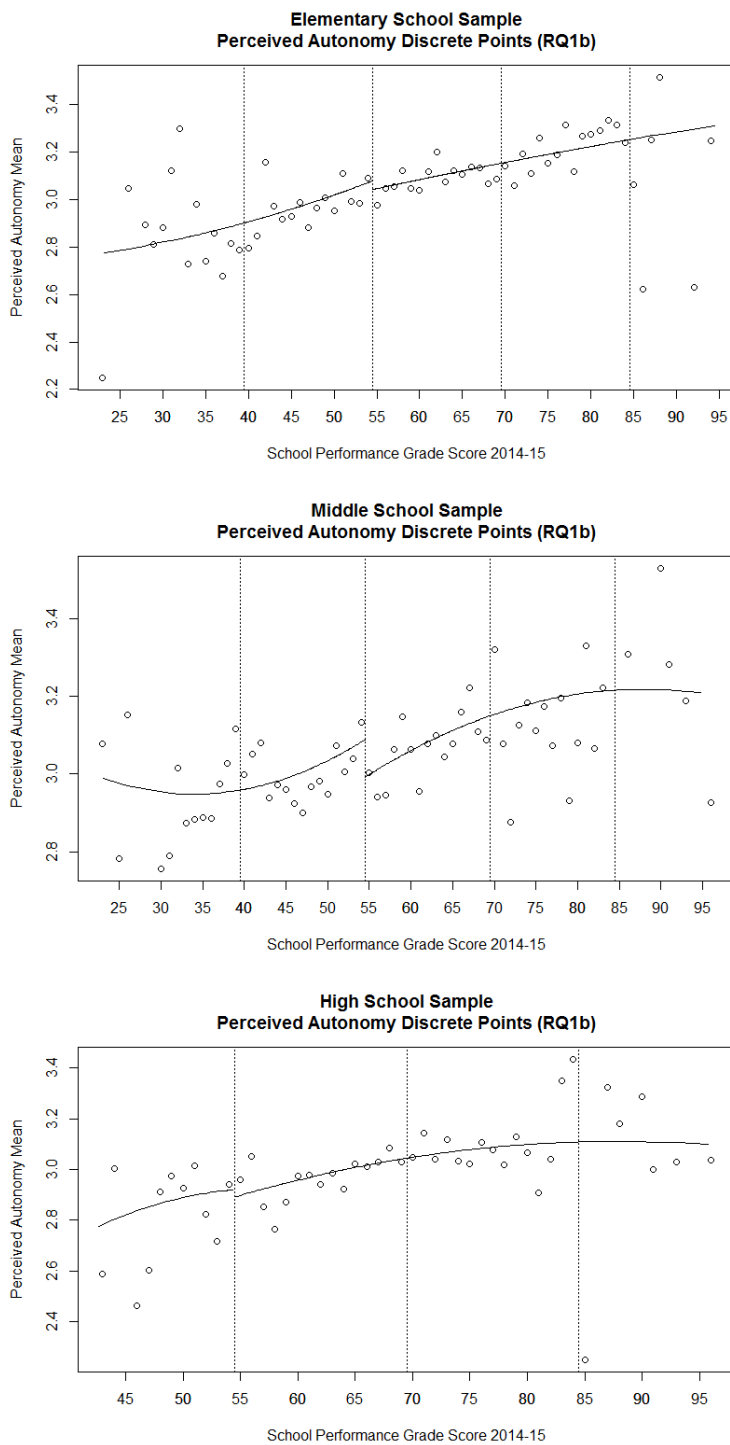


Figure 4.5. Discrete point plots and local pass/fail regressions for perceived autonomy.

*Elementary school sample.* Shown in Figure 4.6, the local regressions for each letter grade showed a potential discontinuity at the D/F threshold, driven mainly by a strong quadratic component in the model for schools receiving an F.

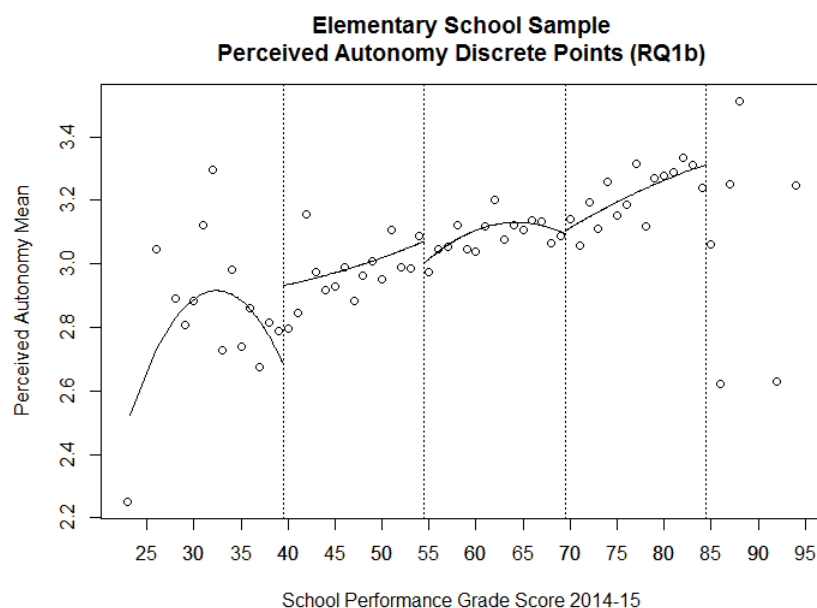


Figure 4.6. Local regressions by letter grade for elementary school perceived autonomy.

Table 4.5 includes the coefficient and standard error values for each model specification, bandwidth, and grade threshold for perceptions of autonomy for the elementary school sample. For all models run, none of the results was statistically significant, including the potential visible discontinuity at the D/F cutoff. Thus, for the elementary school sample, it appears that the School Performance Grade label had no impact on teacher perceptions of autonomy.

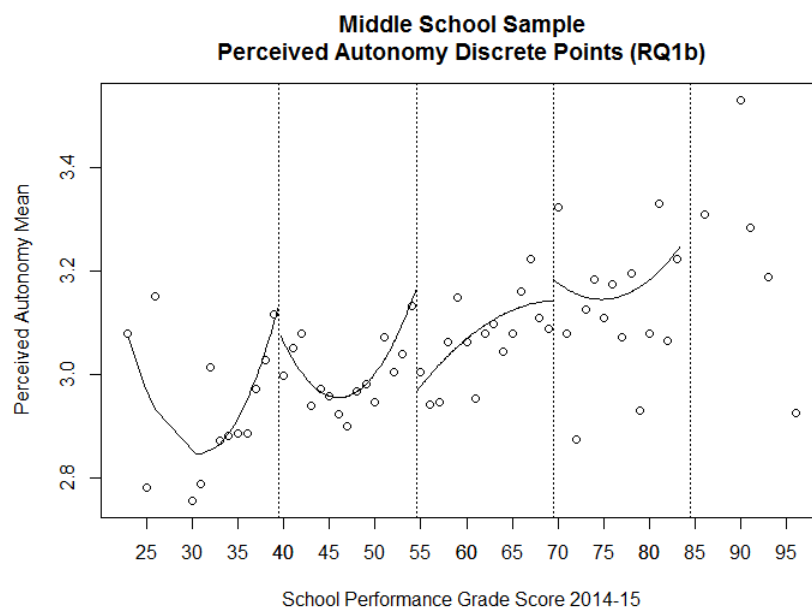
Table 4.5  
*Perceptions of Autonomy – Elementary School Sample*

Threshold	Bandwidth	Sample Size	Rectangular Kernel		Triangular Kernel	
			Linear	Quadratic	Linear	Quadratic
Pass/Fail	All Values	36,390	-0.00	-0.01	-0.00	-0.01
		[1,193]	(0.04)	(0.06)	(0.04)	(0.06)
B/C	10 pt	17,092	-0.03	-0.06	-0.04	-0.03
		[562]	(0.06)	(0.09)	(0.06)	(0.09)
C/D	15 pt	25,036	0.02	-0.04	0.01	-0.04
		[805]	(0.05)	(0.08)	(0.05)	(0.07)
D/F	10 pt	17,149	0.00	0.02	0.01	0.01
		[567]	(0.06)	(0.10)	(0.06)	(0.09)
D/F	15 pt	24,487	-0.04	0.04	-0.04	0.06
		[816]	(0.06)	(0.08)	(0.05)	(0.08)
D/F	10 pt	6,418	-0.19	-0.01	-0.16	0.01
		[226]	(0.12)	(0.18)	(0.12)	(0.17)
D/F	15 pt	10,429	-0.15	-0.14	-0.15	-0.13
		[359]	(0.10)	(0.14)	(0.10)	(0.14)

\* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$ ; Critical values unadjusted for multiple comparisons.

For each analysis, the teacher-level sample size is followed by the school-level sample size in brackets. Each table entry represents the impact estimate of the discontinuity with the standard error of the estimate in parentheses. The entries represent variations in the smoothing kernel and polynomial specification of the running variable, letter grade thresholds, and RD bandwidths.

***Middle school sample.*** The local regressions by letter grade for middle schools, depicted in Figure 4.7, showed a potential discontinuity at the C/D cutoff in the opposite direction from what was expected; observations close to the threshold for D schools indicated higher levels of autonomy than those on the C side. Like the regressions for perceived support, the quadratic term dominated the fit for each letter grade.



*Figure 4.7.* Local regressions by letter grade for middle school perceived autonomy.

Table 4.6 includes the coefficient and standard error values for each model specification, bandwidth, and grade threshold for perceptions of autonomy for the middle school sample. The results contained positive coefficients for each specification and cutoff. One specification at the B/C cutoff ( $0.18 SD$ ,  $p = .046$ ) and one at the C/D cutoff ( $0.16 SD$ ,  $p = .046$ ) showed a *positive* impact of a lower grade on teacher perceptions of autonomy. In other words, taking these results in isolation would indicate receiving a C instead of a B or a D instead of a C led to greater feelings of autonomy. After applying the Benjamini-Hochberg correction across all 28 middle school analyses for autonomy, however, this result was no longer statistically significant. Looking at the evidence in total, it appears there was not an impact of the School Performance Grade label on middle school teachers' perceptions of autonomy.

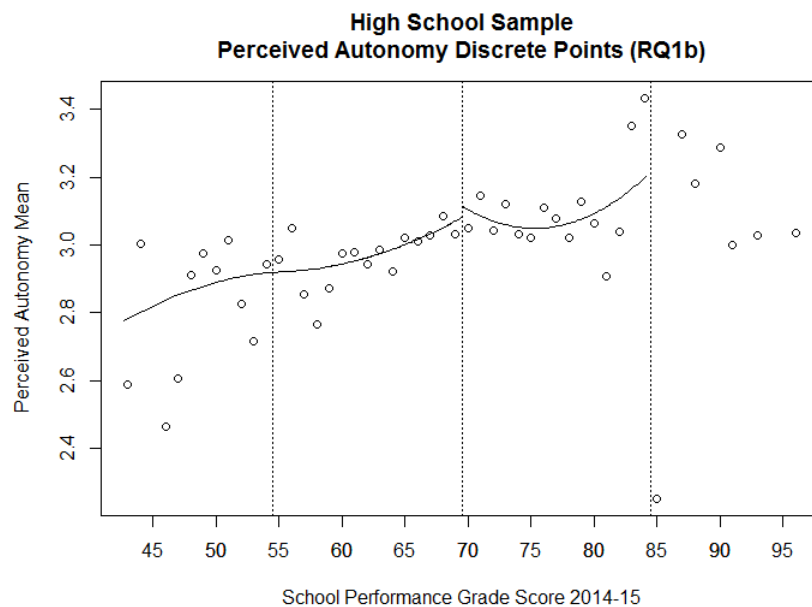
Table 4.6  
*Perceptions of Autonomy – Middle School Sample*

Threshold	Bandwidth	Sample Size	Rectangular Kernel		Triangular Kernel	
			Linear	Quadratic	Linear	Quadratic
Pass/Fail	All Values	15,875 [443]	0.07 (0.06)	0.07 (0.08)	0.07 (0.06)	0.07 (0.08)
		6,600 [177]	0.23 (0.12)	0.16 (0.18)	0.22 (0.12)	0.10 (0.19)
B/C	15 pt	9,501 [256]	<b>0.18*</b> (0.09)	0.21 (0.14)	0.18 (0.09)	0.20 (0.14)
		8,338 [235]	0.16 (0.09)	0.15 (0.13)	<b>0.16*</b> (0.08)	0.17 (0.12)
C/D	15 pt	11,898 [336]	0.10 (0.07)	0.17 (0.11)	0.11 (0.07)	0.18 (0.10)
		3,474 [105]	0.07 (0.16)	-0.00 (0.25)	0.07 (0.16)	0.01 (0.24)
D/F	15 pt	5,724 [169]	0.08 (0.14)	0.13 (0.19)	0.12 (0.14)	0.13 (0.19)

\* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$ ; Critical values unadjusted for multiple comparisons.

For each analysis, the teacher-level sample size is followed by the school-level sample size in brackets. Each table entry represents the impact estimate of the discontinuity with the standard error of the estimate in parentheses. The entries represent variations in the smoothing kernel and polynomial specification of the running variable, letter grade thresholds, and RD bandwidths.

**High school sample.** The high school sample scatter plot indicated a weak positive relationship between a school's continuous School Performance Grade score and the school-level mean for perceived autonomy, consistent with the relationships for the elementary and middle school samples. The local regressions in Figure 4.8 showed no obvious discontinuities at any threshold.



*Figure 4.8.* Local regressions by letter grade for high school perceived autonomy.

Table 4.7 includes the coefficient and standard error values for each model specification, bandwidth, and grade threshold for perceptions of autonomy for the high school sample. For all models run, none of the results was statistically significant at any grade threshold, bandwidth, or model specification. Thus, for the high school sample, it appears that the School Performance Grade label had no impact on teacher perceptions of autonomy.

Table 4.7  
*Perceptions of Autonomy – High School Sample*

Threshold	Bandwidth	Sample Size	Rectangular Kernel		Triangular Kernel	
			Linear	Quadratic	Linear	Quadratic
Pass/Fail	All Values	17,745	-0.02	-0.03	-0.02	-0.04
		[327]	(0.09)	(0.13)	(0.09)	(0.12)
B/C	10 pt	11,938	0.04	-0.00	0.04	-0.01
		[213]	(0.07)	(0.11)	(0.07)	(0.10)
	15 pt	15,385	0.00	0.04	0.01	0.05
		[274]	(0.07)	(0.09)	(0.06)	(0.09)
C/D	10 pt	6,814	-0.02	-0.27	-0.07	-0.25
		[137]	(0.11)	(0.15)	(0.11)	(0.15)
	15 pt	10,629	-0.01	-0.10	-0.02	-0.12
		[205]	(0.09)	(0.13)	(0.09)	(0.13)

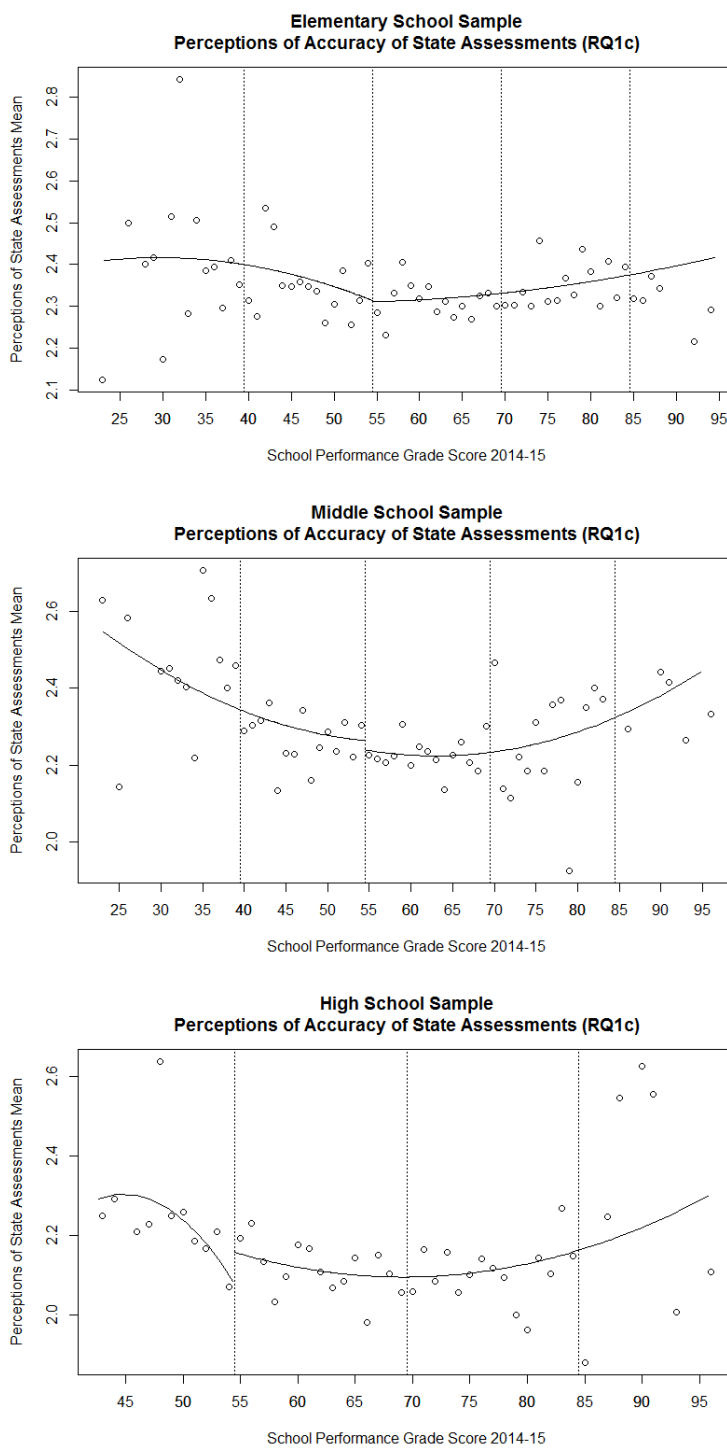
\* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$ ; Critical values unadjusted for multiple comparisons.

For each analysis, the teacher-level sample size is followed by the school-level sample size in brackets. Each table entry represents the impact estimate of the discontinuity with the standard error of the estimate in parentheses. The entries represent variations in the smoothing kernel and polynomial specification of the running variable, letter grade thresholds, and RD bandwidths.

**Perceived accuracy of state assessments.** The plots of discrete points of the school-level means on the item asking teachers about the accuracy of state tests' assessment of student understanding, shown in Figure 4.9, demonstrated no clear linear relationship between the two variables. This relationship was consistent for all three school samples, a surprising finding given my expectation that teachers in higher-performing schools would have a consistently higher perception of the accuracy of the assessments. The means for this item were also significantly lower than for the other measured working conditions constructs in the study. In addition, the plots showed a higher degree of variation than the support and autonomy scales, particularly for the elementary and middle school samples; this was likely due to greater variation of a single item instead of three-item scales.

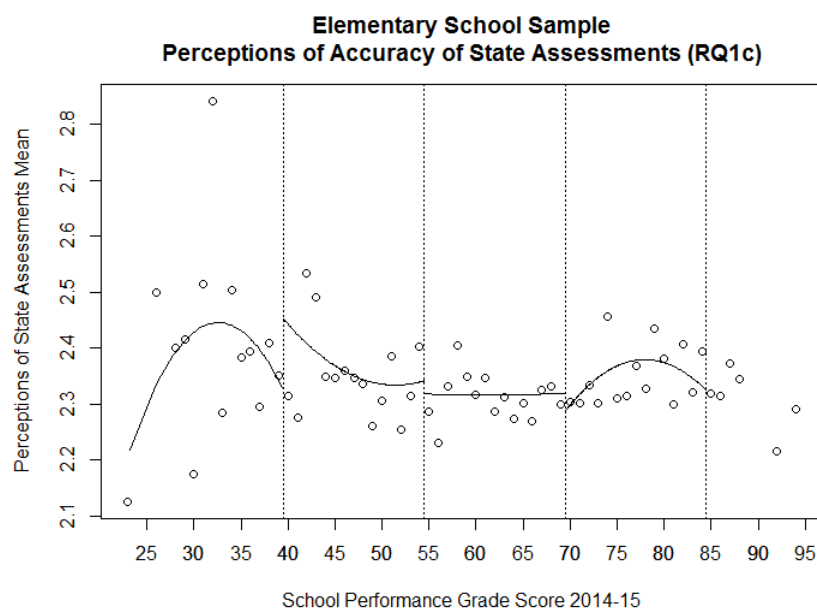
Visual inspection of the plot of pass/fail regressions for elementary schools showed no clear discontinuities for the elementary school sample, but a potential small *negative* discontinuity for the middle school sample and a potential *positive* discontinuity for the high

school sample. The high school regression for schools receiving failing letter grades, however, demonstrated some evidence of overfitting with a strong quadratic component in the curve.



*Figure 4.9.* Discrete point plots and local pass/fail regressions for perceived accuracy of state assessments.

*Elementary school sample.* In Figure 4.10, the local regressions for individual letter grades showed a possible discontinuity between elementary schools at the D/F cutoff, appearing to be driven by a strong quadratic component in the model for schools receiving an F.



*Figure 4.10.* Local regressions by letter grade for elementary school perceptions of state assessment accuracy.

Table 4.8 includes the coefficient and standard error values for each model specification, bandwidth, and grade threshold for perceptions of the accuracy of state assessments for the elementary school sample. For all models run, none of the results were statistically significant, including the D/F cutoff. Thus, for the elementary school sample, it appears that the School Performance Grade label had no impact on teacher perceptions of the accuracy of state assessments.

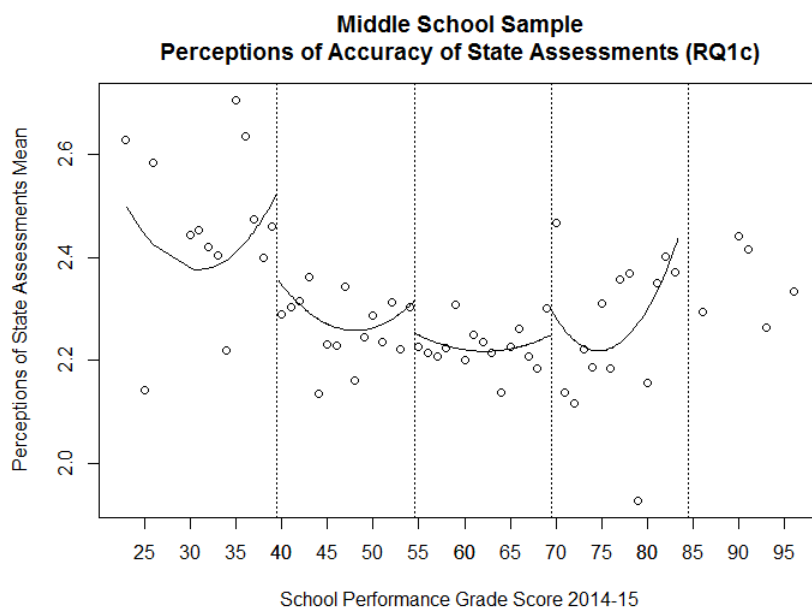
Table 4.8  
*Perceptions of State Assessment Accuracy – Elementary School Sample*

Threshold	Bandwidth	Sample Size	Rectangular Kernel		Triangular Kernel	
			Linear	Quadratic	Linear	Quadratic
Pass/Fail	All Values	35,067	-0.02	-0.03	-0.02	-0.03
		[1,193]	(0.03)	(0.04)	(0.03)	(0.04)
B/C	10 pt	16,523	-0.03	-0.01	-0.03	0.00
		[562]	(0.04)	(0.06)	(0.04)	(0.06)
	15 pt	24,100	-0.02	-0.03	-0.02	-0.01
		[805]	(0.03)	(0.05)	(0.03)	(0.05)
C/D	10 pt	16,506	-0.02	0.11	0.00	0.12
		[567]	(0.04)	(0.07)	(0.04)	(0.06)
	15 pt	23,599	-0.04	0.02	-0.04	0.06
		[816]	(0.04)	(0.05)	(0.04)	(0.05)
D/F	10 pt	6,204	-0.03	0.12	0.02	0.08
		[226]	(0.08)	(0.11)	(0.07)	(0.10)
	15 pt	10,080	0.01	0.01	0.01	0.04
		[359]	(0.06)	(0.09)	(0.06)	(0.09)

\* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$ ; Critical values unadjusted for multiple comparisons.

For each analysis, the teacher-level sample size is followed by the school-level sample size in brackets. Each table entry represents the impact estimate of the discontinuity with the standard error of the estimate in parentheses. The entries represent variations in the smoothing kernel and polynomial specification of the running variable, letter grade thresholds, and RD bandwidths.

***Middle school sample.*** The local regressions in Figure 4.11 showed a potential discontinuity at the D/F threshold, but in a direction opposite of the hypothesis. Teachers on the low end of the cutoff in schools receiving an F indicated a *higher* perceived accuracy of state assessments than teachers on the high end of the cutoff in schools receiving a D. Each of the grade thresholds also showed a potential discontinuity. Like the other local regression curves for individual letter grades in this sample, however, the large influence of the quadratic term indicated potential overfitting to the data that could drive the visual discontinuity.



*Figure 4.11.* Local regressions by letter grade for middle school perceptions of state assessment accuracy.

Table 4.9 contains the coefficients and standard errors for the different grade thresholds, bandwidths, and model specifications. Three models demonstrated statistical significance – both quadratic model at the B/C threshold with a 15-point bandwidth and the linear model with a triangular kernel with a 15-point bandwidth at the D/F threshold. The results do not indicate a consistent pattern, with negative coefficients at the B/C threshold ( $-0.23 SD, p = .044$ ;  $-0.24 SD, p = .031$ ) in the hypothesized direction and a positive coefficient at the D/F threshold ( $0.19 SD, p = .029$ ) in the opposite direction of the hypothesis. After applying the Benjamini-Hochberg correction across all 28 middle school analyses, none of the results demonstrated statistical significance, providing further evidence that these results were due to Type I error. Thus, it appears there was not an impact of the School Performance Grade label on middle school teachers' perceptions of the accuracy of state assessments.

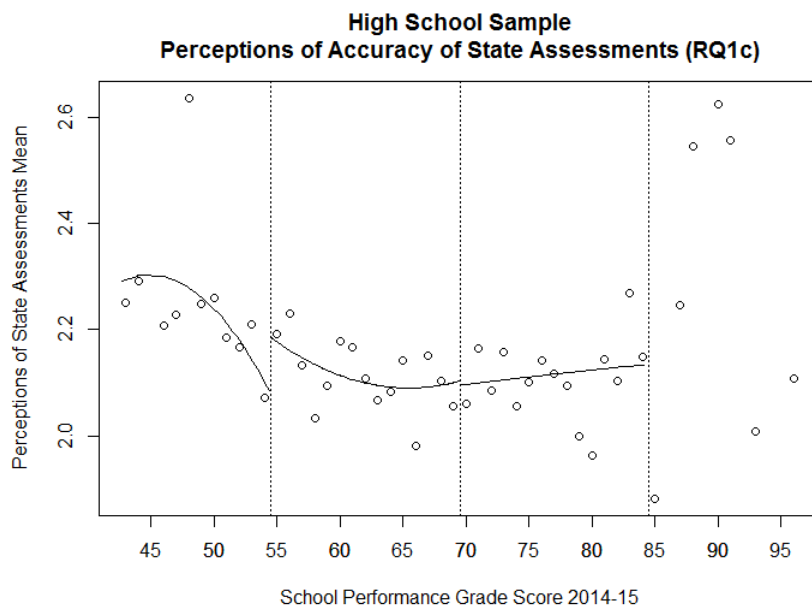
Table 4.9  
*Perceptions of State Assessment Accuracy – Middle School Sample*

Threshold	Bandwidth	Sample Size	Rectangular Kernel		Triangular Kernel	
			Linear	Quadratic	Linear	Quadratic
Pass/Fail	All Values	15,355	-0.05	-0.05	-0.05	-0.05
		[443]	(0.04)	(0.05)	(0.04)	(0.05)
B/C	10 pt	6,348	0.03	<b>-0.23*</b>	-0.04	<b>-0.24*</b>
		[177]	(0.07)	<b>(0.11)</b>	(0.07)	<b>(0.11)</b>
	15 pt	9,168	0.04	-0.03	0.03	-0.09
		[256]	(0.06)	(0.08)	(0.06)	(0.09)
C/D	10 pt	8,086	0.01	-0.06	-0.01	-0.03
		[235]	(0.06)	(0.09)	(0.05)	(0.08)
	15 pt	11,530	-0.04	0.03	-0.02	0.01
		[336]	(0.05)	(0.07)	(0.05)	(0.06)
D/F	10 pt	3,373	0.17	0.12	0.16	-0.01
		[105]	(0.11)	(0.16)	(0.10)	(0.14)
	15 pt	5,562	0.17	0.18	<b>0.19*</b>	0.16
		[169]	(0.09)	(0.13)	<b>(0.09)</b>	(0.12)

\* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$ ; Critical values unadjusted for multiple comparisons.

For each analysis, the teacher-level sample size is followed by the school-level sample size in brackets. Each table entry represents the impact estimate of the discontinuity with the standard error of the estimate in parentheses. The entries represent variations in the smoothing kernel and polynomial specification of the running variable, letter grade thresholds, and RD bandwidths.

**High school sample.** The local regressions in Figure 4.12 showed a possible discontinuity at the C/D threshold, the lowest grade cutoff for the high school sample. Unlike the middle school analysis, this potential discontinuity was in the hypothesized direction with teachers in schools at the low end of the C/D cutoff reporting a lower degree of perceived assessment accuracy than the teachers on the high end of the threshold.



*Figure 4.12.* Local regressions by letter grade for high school perceptions of state assessment accuracy.

Table 4.10 contains the model coefficients and standard errors for each of the grade thresholds, bandwidths, and model specifications. Two of four negative significant coefficients at the Pass/Fail threshold and at the C/D threshold for five of eight specifications suggested a potential negative impact of the receipt of a lower grade on high school teachers' perceptions of the accuracy of state assessments gauging student understanding. Analysis of the B/C threshold indicated no significant impacts.

Table 4.10  
*Perceptions of State Assessment Accuracy – High School Sample*

Threshold	Bandwidth	Sample Size	Rectangular Kernel		Triangular Kernel	
			Linear	Quadratic	Linear	Quadratic
Pass/Fail	All Values	17,040	-0.08	<b>-0.16*</b>	-0.08	<b>-0.16*</b>
		[327]	(0.06)	<b>(0.08)</b>	(0.06)	<b>(0.07)</b>
B/C	10 pt	11,432	-0.01	-0.02	-0.01	-0.03
		[213]	(0.05)	(0.07)	(0.05)	(0.07)
	15 pt	14,762	-0.00	-0.00	-0.01	-0.01
		[274]	(0.04)	(0.06)	(0.04)	(0.06)
C/D	10 pt	6,597	-0.12	<b>-0.29**</b>	<b>-0.16*</b>	<b>-0.31***</b>
		[137]	(0.07)	<b>(0.09)</b>	<b>(0.07)</b>	<b>(0.08)</b>
	15 pt	10,288	-0.10	<b>-0.19*</b>	-0.10	<b>-0.21**</b>
		[205]	(0.06)	<b>(0.08)</b>	(0.06)	<b>(0.08)</b>

\* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$ ; Critical values unadjusted for multiple comparisons.

For each analysis, the teacher-level sample size is followed by the school-level sample size in brackets. Each table entry represents the impact estimate of the discontinuity with the standard error of the estimate in parentheses. The entries represent variations in the smoothing kernel and polynomial specification of the running variable, letter grade thresholds, and RD bandwidths.

Table 4.11 shows the  $p$  values for each of the significant results with application of the Benjamini-Hochberg adjustment. When adjusting for multiple comparisons, these models demonstrated evidence of an impact of the School Performance Grade label on high school teachers at the C/D threshold for four of the eight specifications. The values of three coefficients both remained statistically significant and had values lower than the MDES determined in the power analysis, with impacts ranging from  $-0.23 SD$  to  $-0.31 SD$ . The sensitivity analysis later in the chapter provides some additional evidence that the findings are consistent with other bandwidths and specifications as well. Thus, the analysis suggests that receiving a grade of D versus a grade of C negatively impacted high school teacher perceptions of the accuracy of state assessments.

Table 4.11  
*p* Values and Multiple Comparison Adjustment for High School Perceived Accuracy of Assessments;  $m = 20$  comparisons

Threshold	Bandwidth	Specification	<i>p</i> Value	Rank	Critical <i>p</i> Value After Adjustment	Significant After Adjustment?
C/D	10	Quadratic, Triangular	.000	1	.003	Yes
C/D	10	Quadratic, Rectangular	.002	2	.005	Yes
C/D	15	Quadratic, Triangular	.006	3	.008	Yes
C/D	10	Linear, Triangular	.015	4	.010	No
C/D	15	Quadratic, Rectangular	.027	5	.013	No
Pass/Fail	All	Quadratic, Triangular	.032	6	.015	No
Pass/Fail	All	Quadratic, Rectangular	.047	7	.018	No

The Benjamini-Hochberg method involved sorting the *p* values from least to greatest, and assigning a ranking, with 1 representing the smallest *p* value. The rank, number of comparisons, and critical value (.05) are used to calculate an adjusted critical *p* value for each model that determines whether a model is still significant after multiple comparisons adjustment.

**Perceptions of schools as good places to work and learn.** The final component of the first research question measured teacher responses to a single item asking the degree to which they agreed their schools were good places to work and learn. The perceptions of schools being good places to work and learn correlated strongly with the School Performance Grade score. None of the three samples showed a visible discontinuity at the pass/fail threshold for this outcome.

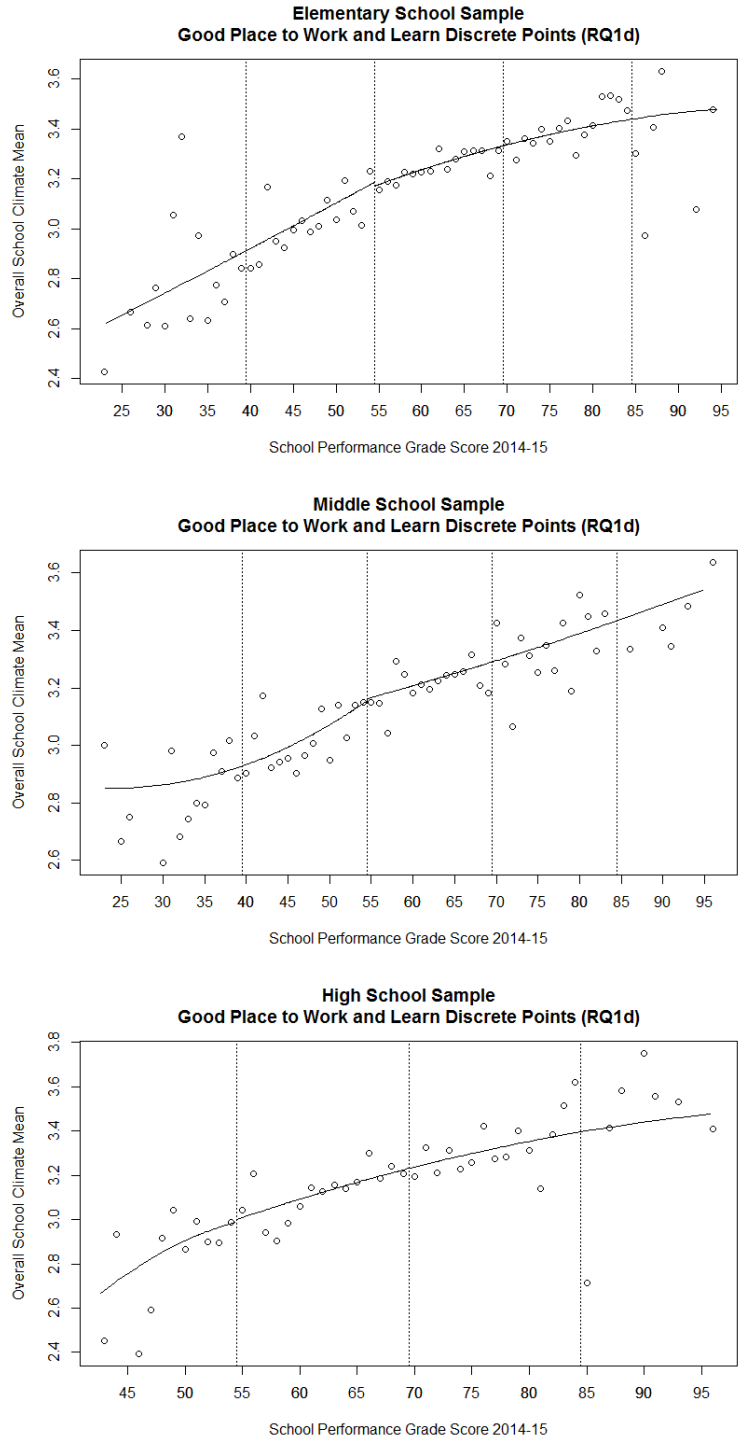
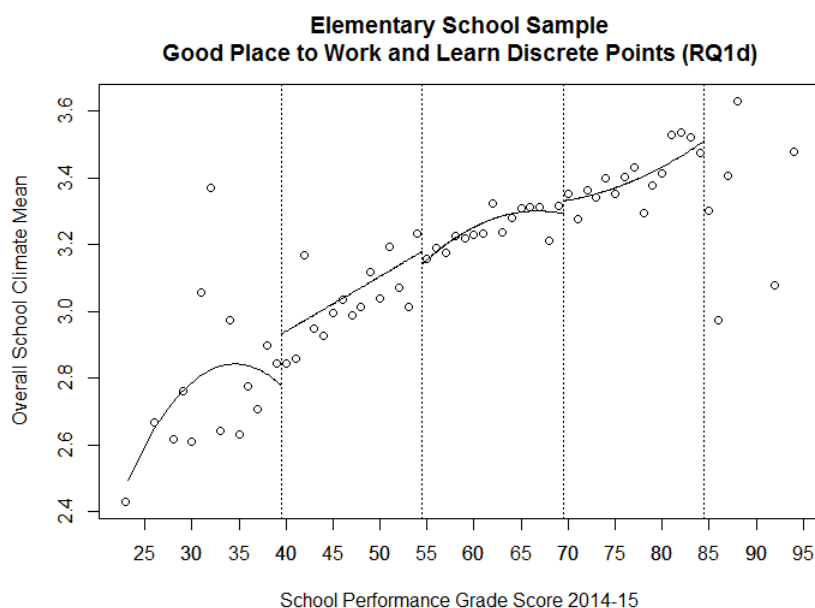


Figure 4.13. Discrete point plots and local pass/fail regressions for perceptions of school as a good place to work and learn.

*Elementary school sample.* The plot of local regressions for each of the letter grades in Figure 4.14 showed a potential discontinuity at the D/F threshold, driven mainly by a strong quadratic component in the model for schools receiving an F.



*Figure 4.14.* Local regressions by letter grade for elementary school perceptions of schools as good places to work and learn.

Table 4.12 summarizes that the impacts of the letter grade were null for each of the thresholds, bandwidths, and model specifications. The potential discontinuity at the D/F threshold that appeared in the plot of discrete values did not demonstrate statistical significance for any of the models. Thus, it appears that the School Performance Grade label had no impact on elementary school teachers' perceptions of their schools as good places to work and learn.

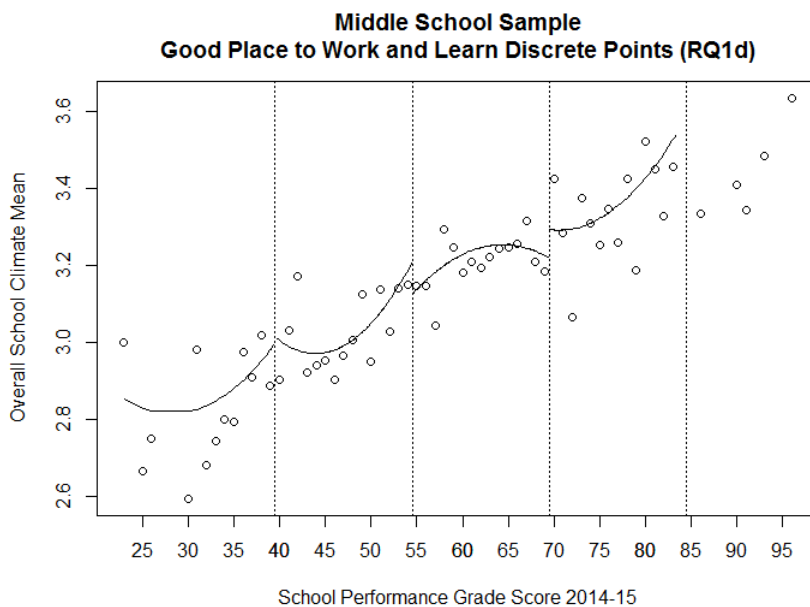
Table 4.12  
*Good Place to Work and Learn – Elementary School Sample*

Threshold	Bandwidth	Sample Size	Rectangular Kernel		Triangular Kernel	
			Linear	Quadratic	Linear	Quadratic
Pass/Fail	All Values	37,015	0.02	-0.01	0.02	-0.01
		[1,193]	(0.04)	(0.05)	(0.04)	(0.05)
B/C	10 pt	17,404	-0.03	-0.01	-0.04	0.01
		[562]	(0.05)	(0.08)	(0.05)	(0.07)
	15 pt	25,463	0.01	-0.05	-0.01	-0.04
		[805]	(0.04)	(0.06)	(0.04)	(0.06)
C/D	10 pt	17,470	0.01	0.07	0.02	0.07
		[567]	(0.06)	(0.09)	(0.06)	(0.08)
	15 pt	24,927	-0.02	0.05	-0.01	0.06
		[816]	(0.05)	(0.07)	(0.05)	(0.07)
D/F	10 pt	6,531	-0.10	0.03	-0.07	0.03
		[226]	(0.11)	(0.16)	(0.11)	(0.16)
	15 pt	10,616	-0.05	-0.10	-0.06	-0.07
		[359]	(0.09)	(0.13)	(0.09)	(0.13)

\* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$ ; Critical values unadjusted for multiple comparisons.

For each analysis, the teacher-level sample size is followed by the school-level sample size in brackets. Each table entry represents the impact estimate of the discontinuity with the standard error of the estimate in parentheses. The entries represent variations in the smoothing kernel and polynomial specification of the running variable, letter grade thresholds, and RD bandwidths.

***Middle school sample.*** The local regressions in Figure 4.15 showed potential discontinuities at the C/D and B/C thresholds, mainly due to a strong quadratic term for schools receiving a C grade.



*Figure 4.15.* Local regressions by letter grade for middle school perceptions of schools as good places to work and learn.

Summarized in Table 4.13, the models showed no significant impacts of the letter grade on middle school teachers' perceptions for all grade thresholds, bandwidths, and model specifications. Thus, it appears there was not an impact of the School Performance Grade label on middle school teachers' perceptions of the accuracy of state assessments as good places to work and learn.

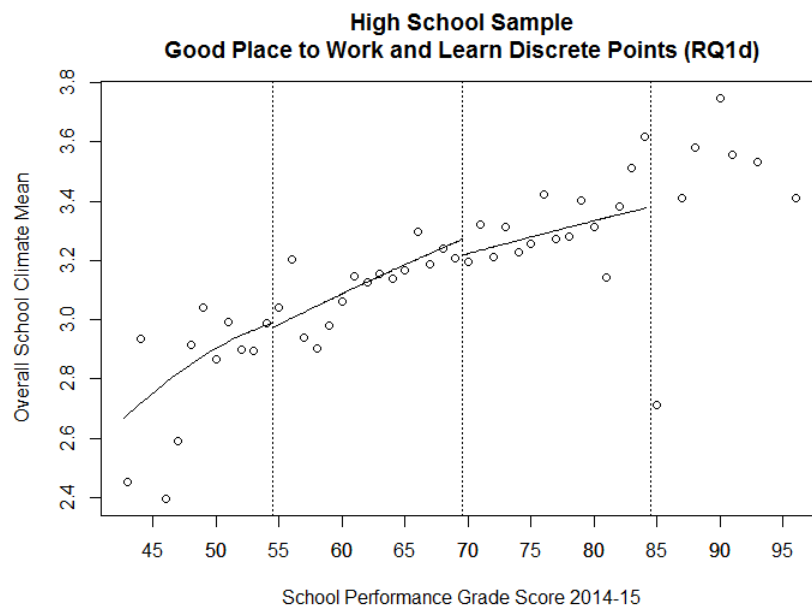
Table 4.13  
*Good Place to Work and Learn – Middle School Sample*

Threshold	Bandwidth	Sample Size	Rectangular Kernel		Triangular Kernel	
			Linear	Quadratic	Linear	Quadratic
Pass/Fail	All Values	16,202	0.01	0.00	0.01	0.00
		[443]	(0.05)	(0.07)	(0.05)	(0.07)
B/C	10 pt	6,743	0.14	0.07	0.14	0.02
		[177]	(0.10)	(0.17)	(0.11)	(0.17)
	15 pt	9,715	0.11	0.13	0.11	0.11
		[256]	(0.08)	(0.13)	(0.08)	(0.13)
C/D	10 pt	8,514	0.04	-0.07	0.03	-0.04
		[235]	(0.08)	(0.12)	(0.08)	(0.11)
	15 pt	12,140	0.01	0.03	0.01	0.04
		[336]	(0.07)	(0.10)	(0.07)	(0.10)
D/F	10 pt	3,519	0.01	0.01	0.02	-0.00
		[105]	(0.15)	(0.23)	(0.15)	(0.23)
	15 pt	5,824	0.06	0.07	0.06	0.07
		[169]	(0.12)	(0.18)	(0.12)	(0.18)

\* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$ ; Critical values unadjusted for multiple comparisons.

For each analysis, the teacher-level sample size is followed by the school-level sample size in brackets. Each table entry represents the impact estimate of the discontinuity with the standard error of the estimate in parentheses. The entries represent variations in the smoothing kernel and polynomial specification of the running variable, letter grade thresholds, and RD bandwidths.

**High school sample.** The plot of discrete points and local regressions in Figure 4.16 demonstrated a strong, positive relationship between the School Performance Grade score and perceptions of schools as a “good place to work and learn” for the high school sample. The plot showed no clear evidence of a discontinuity at the C/D threshold, but a potential small discontinuity in the opposite from expected direction at the B/C threshold.



*Figure 4.16.* Local regressions by letter grade for high school perceptions of schools as good places to work and learn.

The coefficients for the treatment variable in Table 4.14 reflected the visual inspection. For all thresholds, bandwidths, and model specifications tested, only one treatment coefficient demonstrated statistical significance. In isolation, a discontinuity at the B/C threshold indicated *greater* feelings of their schools as good places to work and learn from high school teachers in schools receiving a C than in schools receiving a B, 0.13 *SD*,  $p = .042$ . However, after adjusting for multiple comparisons, this result was no longer statistically significant. Thus, for the high school sample, it appears that the letter grade a school received did not impact high school teachers' perceptions of their schools as good places to work and learn.

Table 4.14  
*Good Place to Work and Learn – High School Sample*

Threshold	Bandwidth	Sample Size	Rectangular Kernel		Triangular Kernel	
			Linear	Quadratic	Linear	Quadratic
Pass/Fail	All Values	18,254	-0.03	-0.03	-0.03	-0.03
		[327]	(0.08)	(0.11)	(0.07)	(0.10)
B/C	10 pt	12,268	<b>0.13*</b>	0.05	0.12	0.01
		[213]	<b>(0.06)</b>	(0.09)	(0.06)	(0.09)
	15 pt	15,820	0.08	0.12	0.10	0.12
		[274]	(0.05)	(0.08)	(0.05)	(0.07)
C/D	10 pt	7,009	-0.01	-0.19	-0.05	-0.13
		[137]	(0.10)	(0.13)	(0.10)	(0.13)
	15 pt	10,931	-0.02	-0.04	-0.02	-0.05
		[205]	(0.08)	(0.11)	(0.08)	(0.11)

\* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$ ; Critical values unadjusted for multiple comparisons.

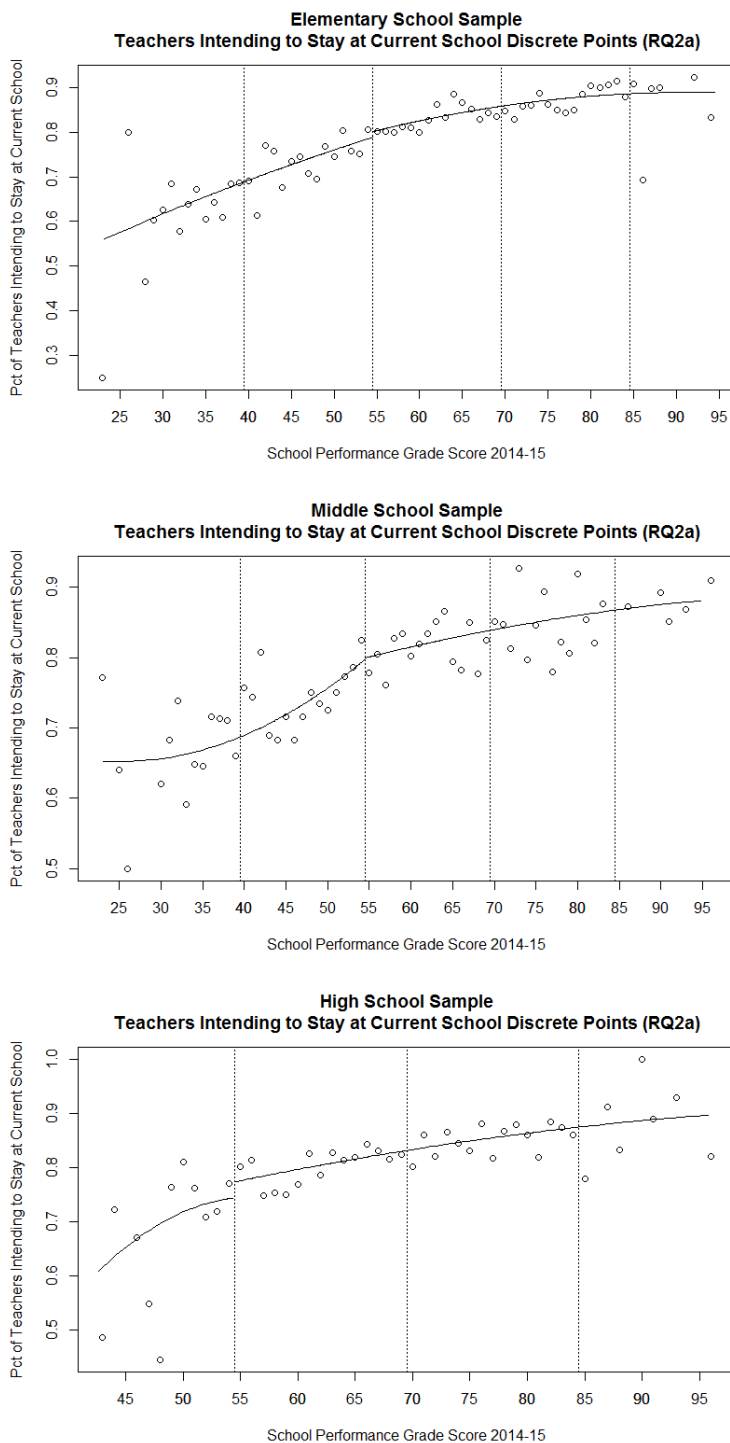
For each analysis, the teacher-level sample size is followed by the school-level sample size in brackets. Each table entry represents the impact estimate of the discontinuity with the standard error of the estimate in parentheses. The entries represent variations in the smoothing kernel and polynomial specification of the running variable, letter grade thresholds, and RD bandwidths.

### Research Question 2: Immediate Professional Plans

The second research question used data from the 2016 Teacher Working Conditions survey to explore teachers' immediate professional plans. The hypothesis derived from the theory of action reasoned that, if the letter grades had an impact, teachers in schools receiving a lower grade would indicate a lower propensity to stay teaching at their current school and a higher propensity to leave education entirely. The data came from a single categorical response in the survey about intent for the next school year and I coded binary outcomes related to a) whether a teacher indicated they would remain teaching at the same school the next year and b) whether a teacher indicated they would leave education entirely in the following school year.

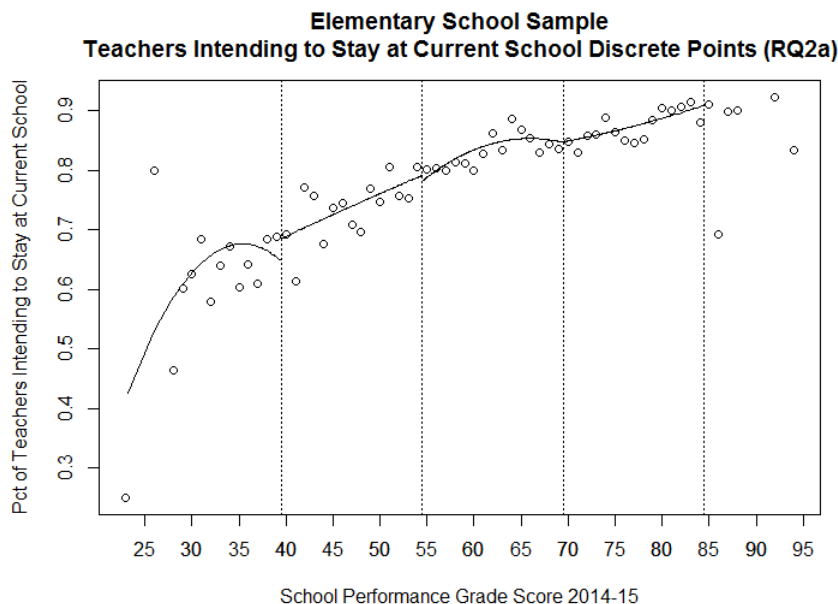
**Intent to remain teaching at the same school next year.** The discrete value plots of the means for the item asking teachers about the proportion of teachers intending to remain teaching at their same school in the following year demonstrated a positive relationship with the School Performance Grade score. Shown in Figure 4.17, this relationship was consistent for all three school samples. The elementary and middle school plots showed no clear discontinuity; the high

school plot of local pass/fail regressions showed a potential small discontinuity in the hypothesized direction between teachers in schools receiving failing vs. passing grades.



*Figure 4.17.* Discrete point plots and local pass/fail regressions for teacher intent to remain teaching at their schools the following year.

*Elementary school sample.* The plot of local regressions for each of the letter grades in Figure 4.18 showed a potential discontinuity at the D/F threshold, driven mainly by a strong quadratic component in the model for schools receiving an F.



*Figure 4.18.* Local regressions by letter grade for elementary school teacher intent to remain teaching at their schools the following year.

Table 4.15 summarizes that all impacts of the letter grade were null for each of the thresholds, bandwidths, and model specifications. The potential discontinuity at the D/F threshold that appeared in the plot of discrete values was not statistically significant for any of the models. Thus, it appears that the School Performance Grade label had no impact on elementary school teachers' intent to remain teaching at their current school.

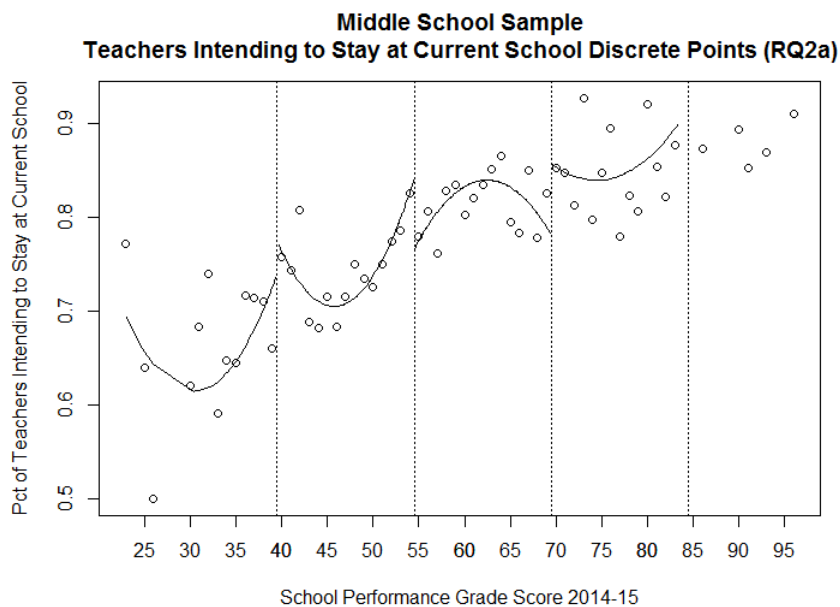
Table 4.15  
*Intent to Stay Teaching at Same School – Elementary School Sample*

Threshold	Bandwidth	Sample Size	Rectangular Kernel		Triangular Kernel	
			Linear	Quadratic	Linear	Quadratic
Pass/Fail	All Values	37,401	-0.00	-0.02	-0.01	-0.01
		[1,193]	(0.02)	(0.02)	(0.02)	(0.02)
B/C	10 pt	17,514	-0.00	-0.03	-0.01	-0.02
		[562]	(0.02)	(0.02)	(0.02)	(0.02)
C/D	15 pt	25,663	0.01	-0.01	0.01	-0.01
		[805]	(0.01)	(0.02)	(0.01)	(0.02)
D/F	10 pt	17,654	0.00	0.00	0.00	0.00
		[567]	(0.02)	(0.03)	(0.02)	(0.03)
D/F	15 pt	25,199	-0.01	0.01	-0.01	0.01
		[816]	(0.02)	(0.02)	(0.01)	(0.02)
D/F	10 pt	6,646	-0.04	-0.00	-0.03	0.01
		[226]	(0.04)	(0.05)	(0.04)	(0.05)
D/F	15 pt	10,796	-0.01	-0.03	-0.01	-0.03
		[359]	(0.03)	(0.04)	(0.03)	(0.04)

\* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$ ; Critical values unadjusted for multiple comparisons.

For each analysis, the teacher-level sample size is followed by the school-level sample size in brackets. Each table entry represents the impact estimate of the discontinuity with the standard error of the estimate in parentheses. The entries represent variations in the smoothing kernel and polynomial specification of the running variable, letter grade thresholds, and RD bandwidths.

***Middle school sample.*** Each of the local regressions depicted in Figure 4.19 contained a visible strong quadratic component, leading to potential discontinuities at the B/C and C/D thresholds. The potential discontinuity at the C/D threshold fell in the opposite direction of the hypothesis, with teachers near the cutoff in schools receiving a D grade indicated a higher propensity to return the following year than teachers near the cutoff in schools receiving a C grade.



*Figure 4.19.* Local regressions by letter grade for middle school teacher intent to remain teaching at their schools the following year.

The results in Table 4.16 show a significant finding for the C/D cutoff for three of the eight models tested. Like the plot of local regressions, these significant coefficients in isolation were opposite the expected direction. However, with only three of eight models showing significance, the findings were not robust to different model parameters.

Table 4.16  
*Intent to Stay Teaching at Same School – Middle School Sample*

Threshold	Bandwidth	Sample Size	Rectangular Kernel		Triangular Kernel	
			Linear	Quadratic	Linear	Quadratic
Pass/Fail	All Values	16,459	-0.00	0.01	-0.00	0.01
		[443]	(0.02)	(0.02)	(0.02)	(0.02)
B/C	10 pt	6,812	-0.02	-0.05	-0.03	-0.05
		[177]	(0.03)	(0.05)	(0.03)	(0.05)
	15 pt	9,824	0.00	-0.03	-0.01	-0.04
		[256]	(0.02)	(0.04)	(0.02)	(0.04)
C/D	10 pt	8,652	0.05	0.05	<b>0.05*</b>	0.06
		[235]	(0.02)	(0.04)	<b>(0.02)</b>	(0.03)
	15 pt	12,339	-0.00	<b>0.08*</b>	0.01	<b>0.07**</b>
		[336]	(0.02)	<b>(0.03)</b>	(0.02)	<b>(0.03)</b>
D/F	10 pt	3,619	-0.05	-0.14	-0.05	-0.16
		[105]	(0.06)	(0.09)	(0.06)	(0.09)
	15 pt	5,963	-0.01	-0.11	-0.02	-0.10
		[169]	(0.05)	(0.06)	(0.05)	(0.07)

\* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$ ; Critical values unadjusted for multiple comparisons.

For each analysis, the teacher-level sample size is followed by the school-level sample size in brackets. Each table entry represents the impact estimate of the discontinuity with the standard error of the estimate in parentheses. The entries represent variations in the smoothing kernel and polynomial specification of the running variable, letter grade thresholds, and RD bandwidths.

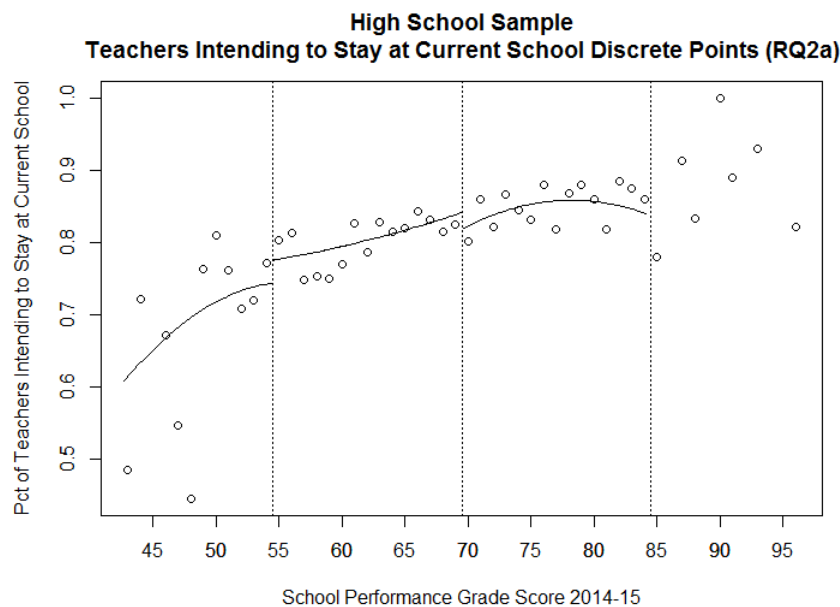
Shown in Table 4.17, application of the Benjamini-Hochberg correction for all the middle school sample comparisons lowered the critical  $p$  values to levels where the results were no longer significant. Thus, it appears that the School Performance Grade had no significant impact on middle school teachers' intent to stay teaching at their same school in the following year.

Table 4.17  
*p* Values and Multiple Comparison Adjustment for Middle School Teacher Intent to Remain Teaching at Their Schools;  $m = 28$  comparisons

Threshold	Bandwidth	Specification	<i>p</i> Value	Rank	Critical <i>p</i> Value After Adjustment	Significant After Adjustment?
C/D	15	Quadratic, Triangular	.009	1	.002	No
C/D	15	Quadratic, Rectangular	.012	2	.004	No
C/D	10	Linear, Triangular	.034	3	.005	No

The Benjamini-Hochberg method involved sorting the *p* values from least to greatest, and assigning a ranking, with 1 representing the smallest *p* value. The rank, number of comparisons, and critical value (.05) are used to calculate an adjusted critical *p* value for each model that determines whether a model is still significant after multiple comparisons adjustment.

**High school sample.** Shown in Figure 4.20, the plot of discrete points for the high school sample showed a positive correlation between the percentage of teachers intending to remain teaching at their schools the next year and the School Performance Grade score a school received. The local regressions indicated a potential discontinuity at the C/D threshold, with teachers in schools at the lowest end of performance of the high school samples particularly indicating a lower level of intent to stay in their schools the next year.



*Figure 4.20.* Local regressions by letter grade for high school teacher intent to remain teaching at their schools the following year.

The analytic results concurred with the appearance of a small discontinuity at the C/D threshold in the expected direction. Table 4.18 contains the results for the high school sample analysis. None of the models for the pass/fail or B/C analysis were statistically significant. Two models at the C/D threshold, with a 10-point bandwidth and quadratic specifications, were independently statistically significant with impacts of nine ( $p = .039$ ) and 10 percentage points ( $p = .016$ ) respectively. Application of the Benjamini-Hochberg correction across all 20 high school sample comparisons, however, lowered the critical  $p$  values to levels where the results were no longer significant. Thus, it appears that the School Performance Grade had no significant impact on high school teachers' intent to stay teaching at their same school in the following year.

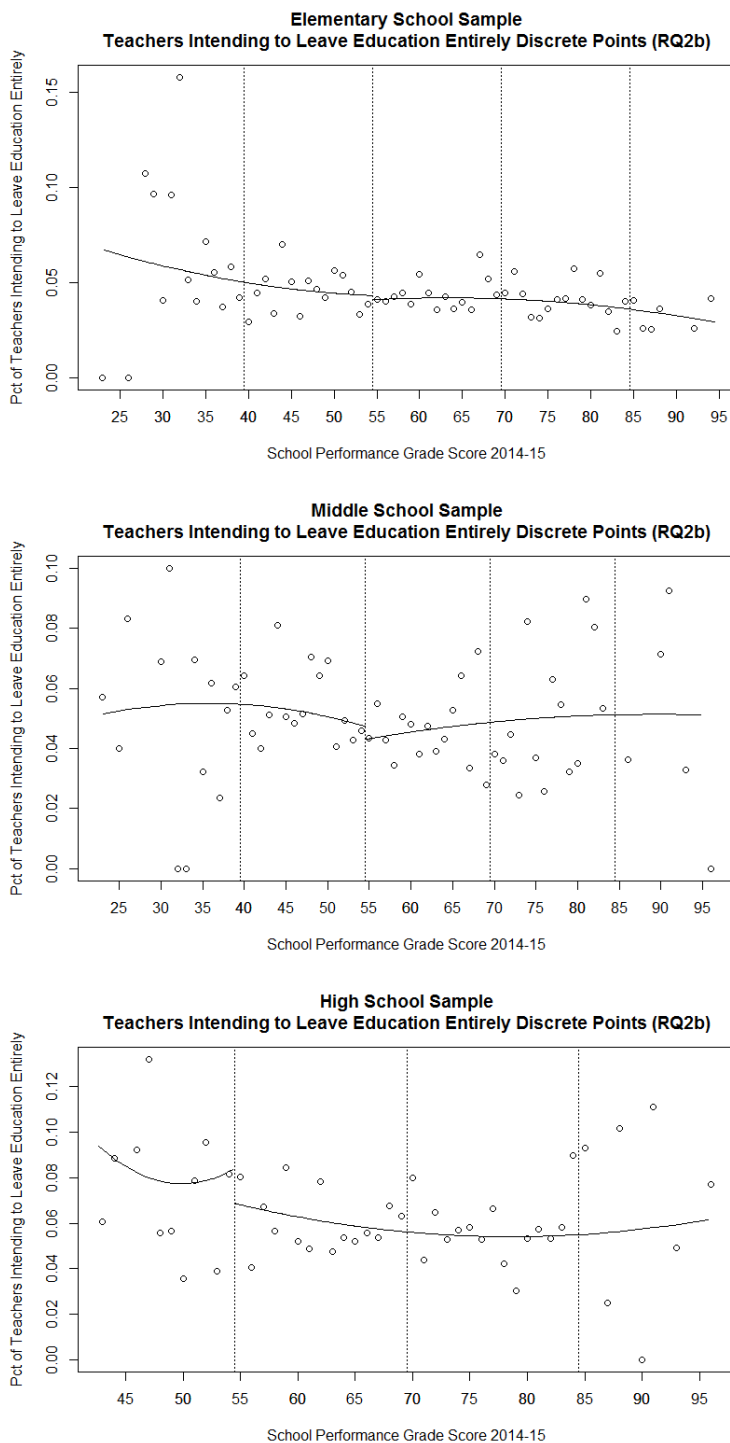
Table 4.18  
*Intent to Stay Teaching at Same School – High School Sample*

Threshold	Bandwidth	Sample Size	Rectangular Kernel		Triangular Kernel	
			Linear	Quadratic	Linear	Quadratic
Pass/Fail	All Values	18,479	-0.02	-0.04	-0.03	-0.04
		[327]	(0.02)	(0.03)	(0.02)	(0.03)
B/C	10 pt	12,405	0.02	0.01	0.02	-0.01
		[213]	(0.01)	(0.02)	(0.01)	(0.02)
	15 pt	16,010	0.01	0.02	0.01	0.02
		[274]	(0.01)	(0.02)	(0.01)	(0.02)
C/D	10 pt	7,125	-0.05	<b>-0.09*</b>	-0.06	<b>-0.10*</b>
		[137]	(0.03)	<b>(0.04)</b>	(0.03)	<b>(0.04)</b>
	15 pt	11,102	-0.03	-0.04	-0.04	-0.06
		[205]	(0.03)	(0.04)	(0.03)	(0.03)

\* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$ ; Critical values unadjusted for multiple comparisons.

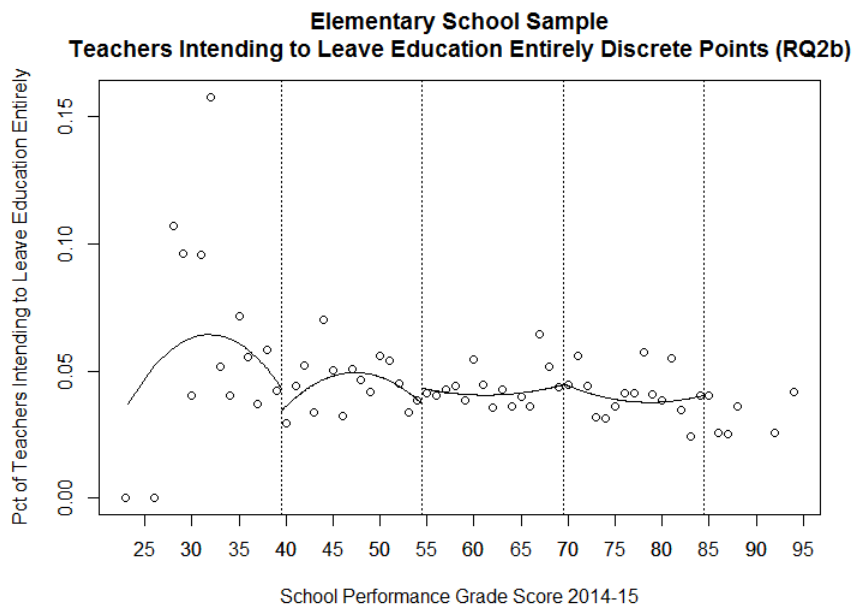
For each analysis, the teacher-level sample size is followed by the school-level sample size in brackets. Each table entry represents the impact estimate of the discontinuity with the standard error of the estimate in parentheses. The entries represent variations in the smoothing kernel and polynomial specification of the running variable, letter grade thresholds, and RD bandwidths.

**Intent to leave education entirely.** The second part of RQ2 explored teachers' intent to leave education entirely. The scatter plots for each sample in Figure 4.21 showed a generally weak negative relationship between the percentage of teachers intending to leave education the next year and the School Performance Grade score. For the elementary school sample, the highest degree of variance occurred within schools that received an F grade. For the middle and high school samples, however, the proportion of teachers intending to leave education demonstrated high variance at the high *and* low ends of the School Performance Grade score distribution. No visible discontinuities appeared at the pass/fail cutoff for the elementary and middle school samples. The high school sample, however, showed a potential discontinuity in the hypothesized direction, with a higher estimate of teachers intending to leave education from schools receiving a failing grade than from schools receiving a passing grade.



*Figure 4.21.* Discrete point plots and local pass/fail regressions for teacher intent to leave teaching entirely.

*Elementary school sample.* The plot of local regressions for each of the letter grades, shown in Figure 4.22, exhibited strong quadratic components for schools receiving grades of D and F and did not show any clear discontinuities at any of the grade thresholds.



*Figure 4.22.* Local regressions by letter grade for elementary school teacher intent to leave education entirely.

Table 4.19 summarizes that all impacts of the letter grade were null for each of the thresholds, bandwidths, and model specifications. Thus, it appears that the School Performance Grade label had no impact on elementary school teachers' intent to leave education entirely.

Table 4.19  
*Intent to Leave Education Entirely – Elementary School Sample*

Threshold	Bandwidth	Sample Size	Rectangular Kernel		Triangular Kernel	
			Linear	Quadratic	Linear	Quadratic
Pass/Fail	All Values	37,401	-0.00	0.00	-0.00	0.00
		[1,193]	(0.00)	(0.01)	(0.00)	(0.01)
B/C	10 pt	17,514	0.00	0.00	0.00	-0.00
		[562]	(0.01)	(0.01)	(0.01)	(0.01)
C/D	15 pt	25,663	0.00	0.00	0.00	0.00
		[805]	(0.01)	(0.01)	(0.01)	(0.01)
D/F	10 pt	17,652	0.00	-0.01	-0.00	-0.01
		[567]	(0.01)	(0.01)	(0.01)	(0.01)
D/F	15 pt	25,199	0.00	-0.00	0.00	-0.00
		[816]	(0.01)	(0.01)	(0.01)	(0.01)
D/F	10 pt	6,646	-0.00	0.02	0.01	0.02
		[226]	(0.01)	(0.02)	(0.01)	(0.02)
D/F	15 pt	10,796	-0.00	0.01	0.00	0.01
		[359]	(0.01)	(0.01)	(0.01)	(0.01)

\* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$ ; Critical values unadjusted for multiple comparisons.

**Middle school sample.** In Figure 4.23, each of the local regressions contained a visible strong quadratic component, showing a potential discontinuity at the B/C threshold.

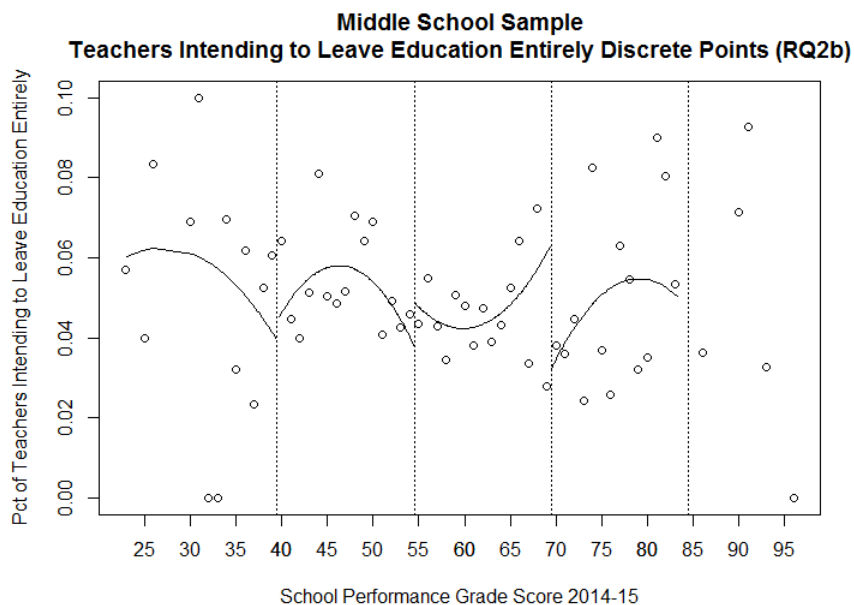


Figure 4.23. Local regressions by letter grade for middle school teacher intent to leave education entirely.

The analysis summarized in Table 4.20 indicated no statistically significant results for any threshold, bandwidth, or model specification. Thus, it appears that the School Performance Grade had no significant impact on middle school teachers' intent to leave education entirely in the following year.

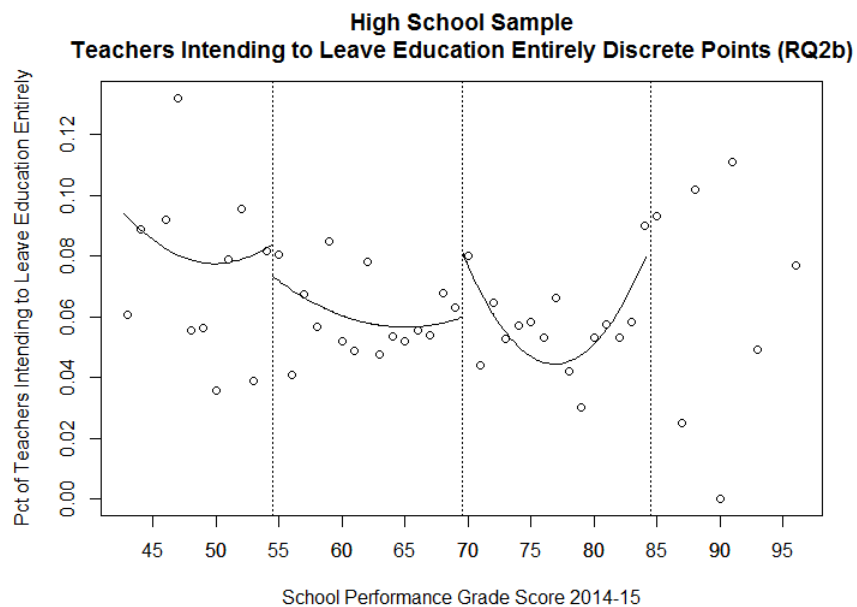
Table 4.20  
*Intent to Leave Education Entirely – Middle School Sample*

Threshold	Bandwidth	Sample Size	Rectangular Kernel		Triangular Kernel	
			Linear	Quadratic	Linear	Quadratic
Pass/Fail	All Values	16,459	0.01	0.00	0.01	0.00
		[443]	(0.01)	(0.01)	(0.01)	(0.01)
B/C	10 pt	6,812	0.02	0.04	0.03	0.03
		[177]	(0.01)	(0.02)	(0.01)	(0.02)
	15 pt	9,824	0.00	0.01	0.03	0.02
		[256]	(0.01)	(0.01)	(0.02)	(0.01)
C/D	10 pt	8,652	0.00	-0.00	-0.00	-0.00
		[235]	(0.01)	(0.01)	(0.01)	(0.01)
	15 pt	12,339	0.01	-0.01	0.00	-0.01
		[336]	(0.01)	(0.01)	(0.01)	(0.01)
D/F	10 pt	3,619	-0.00	0.03	0.01	0.04
		[105]	(0.02)	(0.03)	(0.02)	(0.03)
	15 pt	5,963	-0.02	0.01	-0.01	0.02
		[169]	(0.02)	(0.02)	(0.01)	(0.02)

\* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$ ; Critical values unadjusted for multiple comparisons.

For each analysis, the teacher-level sample size is followed by the school-level sample size in brackets. Each table entry represents the impact estimate of the discontinuity with the standard error of the estimate in parentheses. The entries represent variations in the smoothing kernel and polynomial specification of the running variable, letter grade thresholds, and RD bandwidths.

**High school sample.** Unlike the middle school sample, high school data in Figure 4.24 showed a weak, negative correlation between the School Performance Grade score and the percentage of teachers indicating plans to leave education entirely. The local regressions showed the potential for discontinuities at the C/D and B/C thresholds.



*Figure 4.24.* Local regressions by letter grade for high school teacher intent to leave education entirely.

The analytic results, summarized in Table 4.21, showed no significant findings for any threshold, bandwidth, or model specification tested. Thus, it appears that the School Performance Grade had no significant impact on high school teachers' intent to leave education entirely in the following year.

Table 4.21  
*Intent to Leave Education Entirely – High School Sample*

Threshold	Bandwidth	Sample Size	Rectangular Kernel		Triangular Kernel	
			Linear	Quadratic	Linear	Quadratic
Pass/Fail	All Values	18,479	0.01	0.01	0.01	0.01
		[327]	(0.01)	(0.02)	(0.01)	(0.01)
B/C	10 pt	12,405	-0.01	-0.01	-0.01	-0.01
		[213]	(0.01)	(0.01)	(0.01)	(0.01)
	15 pt	16,010	-0.00	-0.02	-0.01	-0.01
		[274]	(0.01)	(0.01)	(0.01)	(0.01)
C/D	10 pt	7,125	0.01	0.02	0.02	0.02
		[137]	(0.01)	(0.02)	(0.02)	(0.02)
	15 pt	11,102	0.02	0.01	0.02	0.02
		[205]	(0.01)	(0.02)	(0.01)	(0.02)

\* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$ ; Critical values unadjusted for multiple comparisons.

For each analysis, the teacher-level sample size is followed by the school-level sample size in brackets. Each table entry represents the impact estimate of the discontinuity with the standard error of the estimate in parentheses. The entries represent variations in the smoothing kernel and polynomial specification of the running variable, letter grade thresholds, and RD bandwidths.

### Research Question 3: Actual Turnover

The analysis for RQ3 used a single-level model that looked at each grade threshold for each sample using the school-level one-year teacher turnover percentage. Unlike the Teacher Working Conditions survey that is administered only every two years, the teacher turnover percentage measure is available for each year. The timing of this measure, however, represented a large amount of lag. The most recently-available teacher turnover percentage in the public file represents the 2016-17 school year. Thus, the most recent data available represent teachers who left their schools prior to or during the 2015-16 school year. Due to the timing of the release of letter grades, the 2013-14 School Performance Grade, released in February 2015, serves as the treatment.

The exploration of RQ3 using an outcome with more immediate timing following the initial assignment of the first School Performance Grades afforded a more unobstructed view of the policy's impact. The timing of the analysis for actual turnover avoided the potential contamination of having multiple School Performance Grades assigned to schools in 2013-14

and 2014-15 between administrations of the NC Teacher Working Conditions survey as was the case with the first two research questions.

In addition to the outcome and treatment variables, each single-level model included various specifications of the running variable (the 2013-14 School Performance Grade score), the teacher turnover percentage from the two previous school years, the percentage of economically disadvantaged students, and the percentage of students in the six race categories.

The plots in Figure 4.25 show a general negative correlation between the rate of teacher turnover and the School Performance Grade scores received by schools. The local pass/fail regressions showed potential small discontinuities in the expected direction.

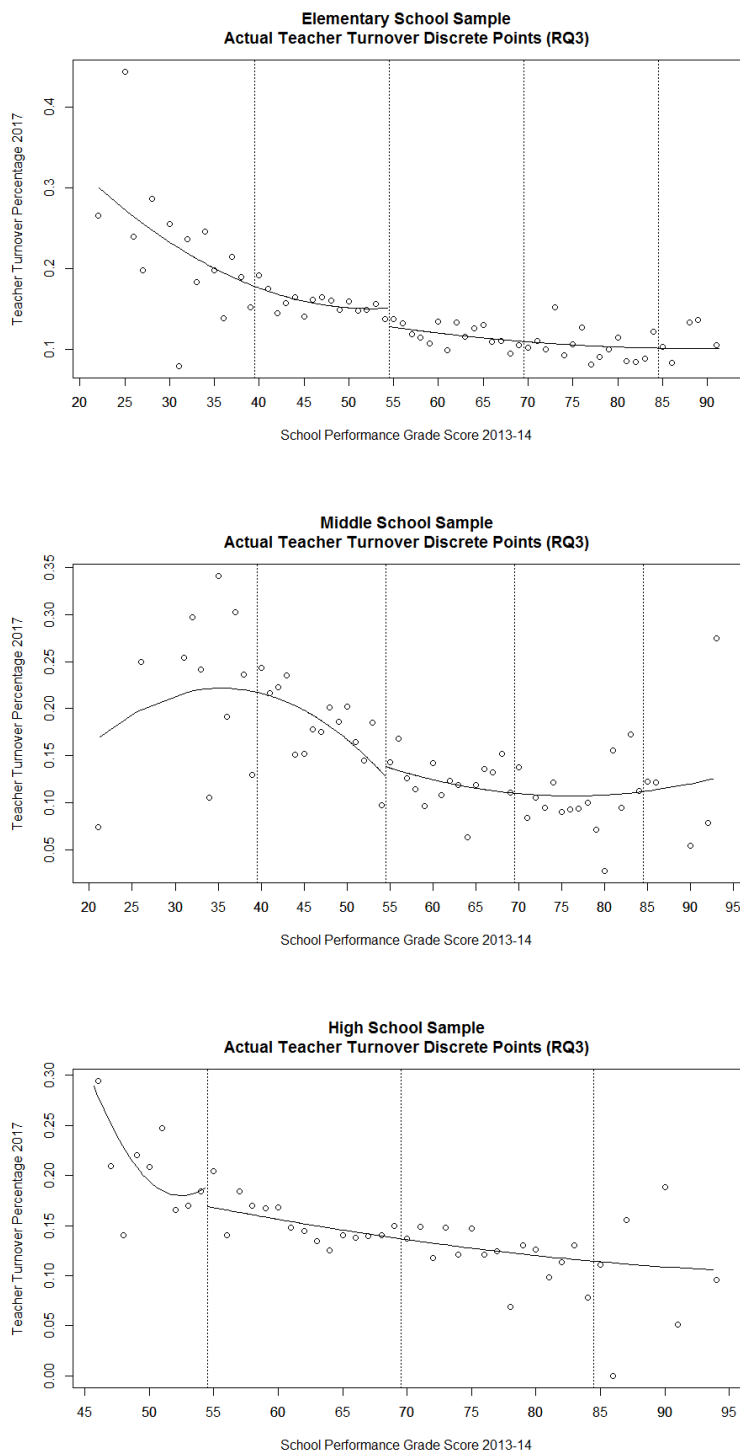
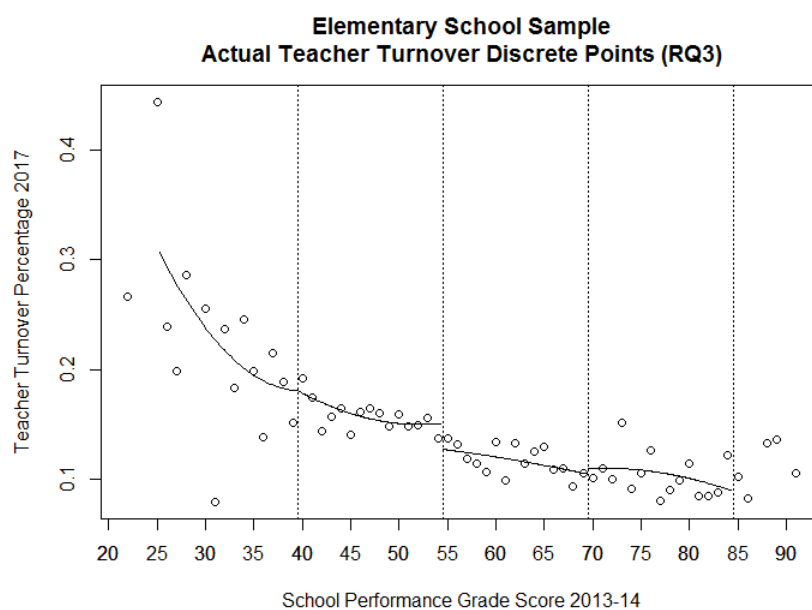


Figure 4.25. Discrete point plots and local pass/fail regressions for actual teacher turnover.

**Elementary school sample.** The plot of school-level teacher turnover in the year following the receipt of the 2013-14 School Performance Grade for elementary schools showed a

negative relationship between the two variables in the expected direction; schools earning a higher School Performance Grade scores had lower rates of teacher turnover. The rate of teacher turnover was disproportionately large in schools receiving an F grade. The local regressions in Figure 4.26 indicated a potential discontinuity at the C/D threshold.



*Figure 4.26.* Local regressions by letter grade for elementary school teacher turnover.

The results displayed in Table 4.22 showed no significant impacts for any grade threshold, bandwidth, or model specification at the Pass/Fail, B/C, and C/D thresholds. One model at the D/F threshold was independently statistically significant in the opposite of the hypothesized direction, with an impact of six percentage points ( $p = .031$ ). Application of the Benjamini-Hochberg correction across all 28 elementary school models for this outcome, however, made this result no longer significant. Thus, it appears that the letter grade that elementary schools received did not significantly impact the teacher turnover rate the following year.

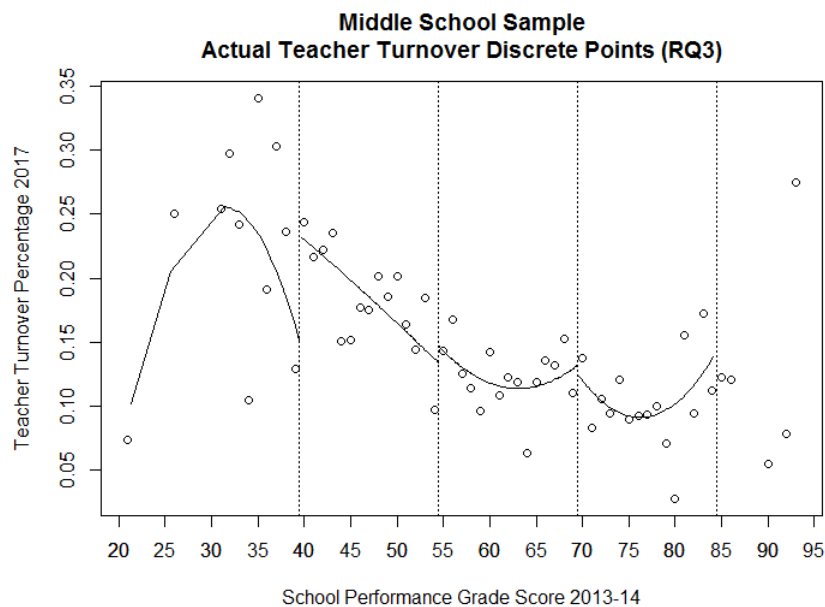
Table 4.22  
*Actual Teacher Turnover – Elementary School Sample*

Threshold	Bandwidth	Sample Size	Rectangular Kernel		Triangular Kernel	
			Linear	Quadratic	Linear	Quadratic
Pass/Fail	All Values	1,193	0.01 (0.01)	0.02 (0.01)	0.01 (0.01)	0.02 (0.01)
	10 pt	547	-0.01 (0.01)	-0.01 (0.02)	-0.01 (0.01)	-0.01 (0.01)
B/C	15 pt	760	-0.01 (0.01)	-0.01 (0.01)	-0.01 (0.01)	-0.01 (0.01)
	10 pt	598	0.02 (0.01)	0.00 (0.02)	0.01 (0.01)	0.00 (0.02)
C/D	15 pt	844	0.01 (0.01)	0.02 (0.01)	0.01 (0.01)	0.01 (0.01)
	10 pt	261	0.00 (0.02)	-0.04 (0.03)	-0.01 (0.02)	<b>-0.06*</b> <b>(0.03)</b>
D/F	15 pt	417	-0.00 (0.02)	-0.00 (0.03)	-0.00 (0.02)	-0.03 (0.02)

\* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$ ; Critical values unadjusted for multiple comparisons.

For each analysis, the teacher-level sample size is followed by the school-level sample size in brackets. Each table entry represents the impact estimate of the discontinuity with the standard error of the estimate in parentheses. The entries represent variations in the smoothing kernel and polynomial specification of the running variable, letter grade thresholds, and RD bandwidths.

***Middle school sample.*** The local regressions in Figure 4.27 indicated that a potential discontinuity existed at the D/F threshold. It should be noted, however, that the regression curve for schools receiving an F grade had a strong quadratic component and the location of the curve at the D/F threshold responded to a large degree of variation among schools receiving F grades. The discontinuity was also not in the expected direction, indicating that at the cutoff, schools receiving an F grade had lower teacher turnover than schools receiving a D grade with similar performance.



*Figure 4.27.* Local regressions by letter grade for middle school teacher turnover.

Table 4.23 includes the coefficients for the treatment variable that estimate the impact of the grade received on its rate of teacher turnover the following year. No values were statistically significant. Thus, it appears that the letter grade that middle schools received did not significantly impact the teacher turnover rate the following year.

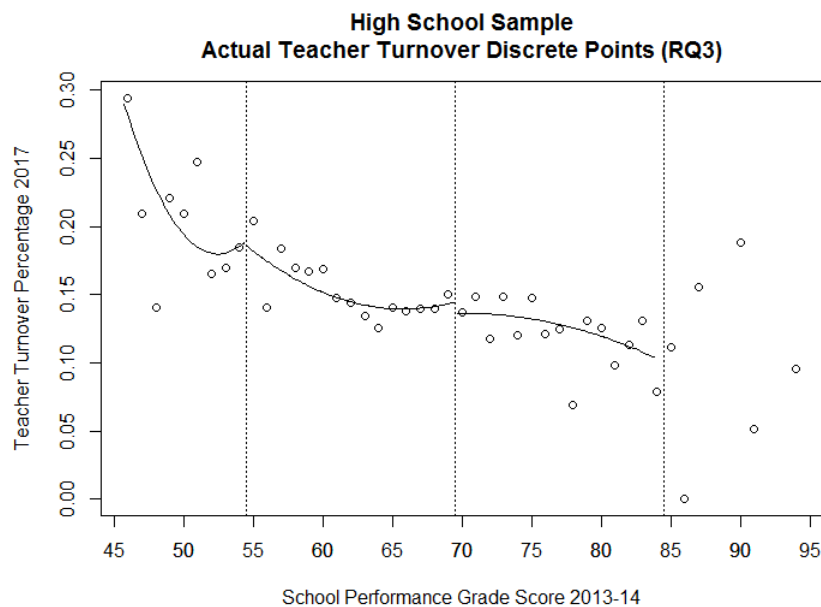
Table 4.23  
*Actual Teacher Turnover – Middle School Sample*

Threshold	Bandwidth	Sample Size	Rectangular Kernel		Triangular Kernel	
			Linear	Quadratic	Linear	Quadratic
Pass/Fail	All Values	446	0.01 (0.01)	-0.01 (0.02)	0.01 (0.01)	-0.01 (0.02)
	10 pt	173	-0.00 (0.02)	0.01 (0.02)	0.00 (0.01)	-0.00 (0.02)
B/C	15 pt	256	0.02 (0.02)	0.01 (0.02)	0.01 (0.01)	-0.00 (0.02)
	10 pt	247	-0.00 (0.02)	-0.03 (0.03)	-0.01 (0.02)	-0.04 (0.02)
C/D	15 pt	340	0.00 (0.02)	0.01 (0.02)	0.00 (0.01)	-0.02 (0.02)
	10 pt	114	-0.04 (0.05)	-0.05 (0.07)	-0.05 (0.04)	-0.07 (0.06)
D/F	15 pt	181	-0.03 (0.03)	-0.03 (0.05)	-0.04 (0.03)	-0.04 (0.05)

\* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$ ; Critical values unadjusted for multiple comparisons.

For each analysis, the teacher-level sample size is followed by the school-level sample size in brackets. Each table entry represents the impact estimate of the discontinuity with the standard error of the estimate in parentheses. The entries represent variations in the smoothing kernel and polynomial specification of the running variable, letter grade thresholds, and RD bandwidths.

**High school sample.** Like the other two samples, the high school rates of teacher turnover displayed a negative relationship with the School Performance Grade score, particularly for schools receiving a C or B grade in 2013-14. The local regressions in Figure 4.28, however, did not indicate clear discontinuities at any of the grade thresholds.



*Figure 4.28.* Local regressions by letter grade for high school teacher turnover.

The analysis summarized in Table 4.24 confirmed the visual inspection that the letter grade a school received did not significantly impact subsequent teacher turnover for any threshold, bandwidth, or model specification. Thus, it appears that the letter grade that high schools received did not significantly impact the teacher turnover rate the following year.

Table 4.24  
*Actual Teacher Turnover – High School Sample*

Threshold	Bandwidth	Sample Size	Rectangular Kernel		Triangular Kernel	
			Linear	Quadratic	Linear	Quadratic
Pass/Fail	All Values	325	-0.02 (0.02)	0.01 (0.03)	-0.01 (0.02)	0.01 (0.03)
	10 pt	216	0.00 (0.01)	0.02 (0.02)	0.01 (0.01)	0.01 (0.02)
B/C	15 pt	289	-0.00 (0.01)	0.01 (0.02)	0.00 (0.01)	0.01 (0.02)
	10 pt	142	-0.02 (0.03)	0.01 (0.04)	0.01 (0.03)	0.02 (0.04)
C/D	15 pt	198	-0.01 (0.03)	-0.00 (0.04)	0.00 (0.02)	0.01 (0.03)

\* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$ ; Critical values unadjusted for multiple comparisons.

For each analysis, the teacher-level sample size is followed by the school-level sample size in brackets. Each table entry represents the impact estimate of the discontinuity with the standard error of the estimate in parentheses. The entries represent variations in the smoothing kernel and polynomial specification of the running variable, letter grade thresholds, and RD bandwidths.

### Sensitivity Analysis to Confirm Statistically Significant Results

Across all analyses, two outcomes were statistically significant after adjusting for multiple comparisons – high school teachers’ perceived support at the C/D threshold and high school teachers’ perception of the accuracy of state assessments at the C/D threshold. For these two outcomes, I ran additional sensitivity analyses.

The main analysis specified two bandwidths with 10 and 15 points of the School Performance Grade cutoffs. I tested three additional bandwidths for each of the significant high school outcomes within five, seven, and 12 points of the School Performance Grade cutoffs as an additional robustness check. Due to the smaller sample sizes for the five and seven point bandwidths, I did not use the quadratic specification or the triangular kernel. Shown in Table 4.25, the results for support demonstrated  $p$  values less than .05 for all of the additional tested specifications. The findings for high school teachers’ perceived support were robust to this additional test for the four additional specifications.

The perception of the accuracy of state assessments was also significant when restricting the bandwidth to 5 and 7 points. Expanding to a 12 point bandwidth did not yielded a statistically significant result for the quadratic specification. In total, there is additional evidence that receiving a grade of D negatively impacted teacher perceptions of state assessment accuracy, but the findings are not as robust to different specifications as those for perceived support.

Table 4.25

*Additional Specifications for Significant Results – High School Sample*

Outcome	Bandwidth at C/D Threshold	Sample Size	Rectangular Kernel	
			Linear	Quadratic
Support	5 pt	3,266 [67]	<b>-0.40**</b> (0.14)	----
	7 pt	4,418 [93]	<b>-0.36**</b> (0.11)	----
	12 pt	8,034 [163]	<b>-0.17*</b> (0.08)	<b>-0.30*</b> (0.12)
Accuracy of State Assessments	5 pt	3,360 [67]	<b>-0.32**</b> (0.10)	----
	7 pt	4,529 [93]	<b>-0.23**</b> (0.08)	----
	12 pt	7,841 [163]	-0.12 (0.06)	<b>-0.23*</b> (0.10)

\* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$ ; Critical values unadjusted for multiple comparisons.

**Summary**

The combination of all analysis suggests that the School Performance Grade label a school received did not have a detectable impact on elementary and middle school teachers for any of the measured constructs. For the high school sample, it appears that receiving a grade of D versus a grade of C made a statistically significant impact on teachers' perceptions of support and their perceptions of the accuracy of state assessments. The final chapter discusses these results in the context of the study conceptual framework, offers directions for future research, suggests mechanisms behind the impact of School Performance Grades and how they could be measured, and provides recommendations and implications for policymakers.

## CHAPTER 5

The analysis showed no consistent significant evidence of the School Performance Grade label a school received impacting teachers in elementary or middle schools. For high schools, however, the results suggested that receiving a D versus a C grade yielded lower responses to the questions related to a) perceived support from parents and the community and b) perceived accuracy of state assessment results accurately reflecting students' understanding of standards. These results remained significant after adjustment of the critical values for multiple comparisons and were robust to sensitivity analysis with varying bandwidths, cut points, and model specifications.

This final chapter begins with interpretation of the results through the lens of the study's theory of action. I then suggest some general directions for future research on A-F school grades. In addition, I provide an analysis of potential explanations for the results that connect to prior research. In cases where an empirical design is possible, I present suggestions for future research to better understand and isolate each potential explanatory mechanism associated with the connection between the A-F grade label and teacher perceptions of their schools. The chapter concludes with implications of the results for policymakers considering introducing and continuing A-F policies in North Carolina and other states.

### **Connecting the Findings to the Theory of Action**

In the theory of action, pictured in Figure 5.1, I hypothesized that the A-F letter grade labels applied to schools in North Carolina could impact teacher perceptions of support, autonomy, accuracy of state assessments, and whether their school was a good place to work and learn. In addition, the framework posited that these working conditions, in combination with the letter grade itself, could impact teacher decisions to remain teaching in their schools or to leave

education. The theory of action suggested two mechanisms for impacting teachers – a) directly through teacher reactions to the stigma of the labels and b) indirectly through changes in the behavior of school leaders, parents, and community members perceived by teachers.

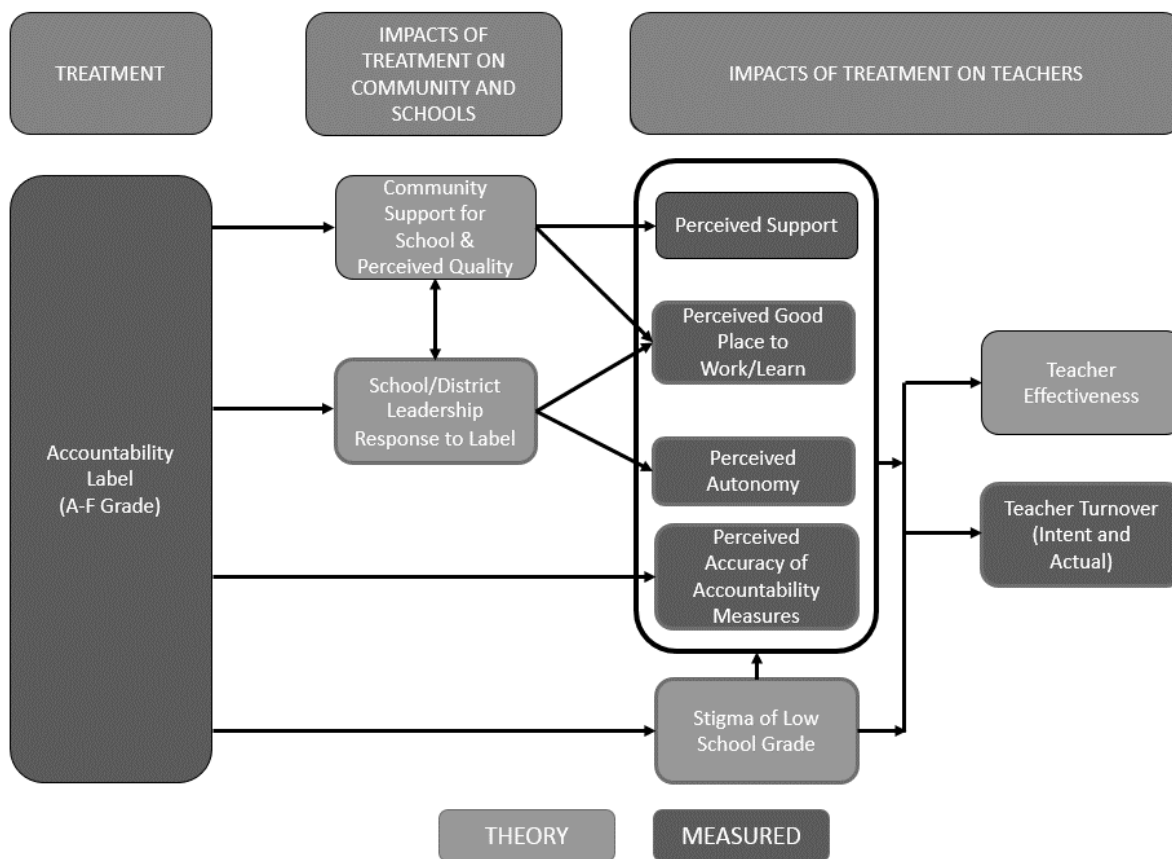


Figure 5.1. Theory of action.

The results suggest that the School Performance Grade labels provided to schools had little *direct* impact on teachers' perceptions of their schools from the stigma of a low grade alone. If this were the case, the models would likely have exhibited an impact on teacher perceptions of their schools as good places to work and learn and influence their decisions to remain teaching in their schools or to leave education entirely. The null results in these domains are consistent with Peterson, Henderson, and West (2014, p. 23) who stated, "On most issues, teacher opinion does not change in response to new information nearly as much as it does for the

public as a whole.” In other words, teachers are less responsive to new information about education than people who work outside of schools.

In addition, the results suggest that the response of school leaders to the grades did not significantly impact teachers. The lack of consistently significant impacts for any sample or letter grade threshold for teacher perceptions of autonomy support this perspective. Changes in autonomy would likely be influenced most closely by school leadership. The null findings for autonomy suggest that teachers did not perceive diminished trust as professionals from their administrators because of the grade their school received.

Perceived support had the most evidence of impact from the letter grade a school received than any of the other constructs measured. Conceptually, this construct is most closely linked to how teachers were affected *indirectly* by others’ responses to the letter grade from outside the school including parents and community members. These groups likely experienced a greater level of information gain from the letter grade than educators working within schools. Chingos, Henderson, and West (2012) noted that accountability systems with easy-to-interpret metrics, such as A-F grades, subsidize the cost of information acquisition by the public. It is possible that people from outside of the schools formed a stronger opinion of their local schools from the introduction of a letter grade in a way that they did not respond to prior forms of school accountability data. This is also consistent with the findings from Peterson, Henderson, and West (2014) in that teachers are less responsive to school performance information than the public. Interpreting the results of my study in the context of the literature makes it possible that people from outside schools responded to the A-F grades in a manner that impacted teachers’ perceptions of their support. Although only the high school sample showed significance in the models, all three samples showed consistently negative coefficients for the pass/fail analysis for

all tested bandwidths and model specifications. In a study with a larger sample of schools or with additional years of data, researchers could conduct a study with the appropriate level of statistical power to detect a potential effect.

The analysis also revealed a significant impact at the C/D threshold for high school teachers on their perception of state assessments accurately reflecting student understanding of the standards. The results for elementary and middle school teachers, however, did not demonstrate significance. Thus, it is possible that high school teachers responded differently to the A-F grade regarding their opinions about state assessment accuracy. As noted in Chapters 1 and 3, unlike elementary and middle schools whose grades are based solely from achievement and growth on end-of-grade tests, high school letter grades are determined by three end-of-course subject tests as well as performance on the ACT and WorkKeys exams and graduation rate.

Also, unlike elementary and middle school teachers who are measured on a larger battery of tests dealing directly with their classroom practice, fewer high school teachers are *directly* measured by the letter grade. Thus, the detection of an effect only for the high school sample could indicate a different orientation toward accountability results for high school teachers who have less connection between their student performance and the assignment of a school's letter grade.

### **Suggested Areas for General Future Research on NC School Performance Grades**

Many areas for further study emerged from this analysis. One potential area for future research involves replicating this study with updated data by testing the same hypotheses and models that yielded significant impacts using the 2016-17 School Performance Grades and results from the recently-released 2018 NC Teacher Working Conditions survey. Pooling the

results from the 2016 and 2018 surveys with their associated school accountability grades would increase statistical power to confirm null impacts for the elementary and middle school samples. At the high school level, the expansion of the sample would provide further evidence to confirm or call into question the individual significant and null findings respectively.

Looking outside of North Carolina could also provide greater insight about the broader impact of A-F school grades on teachers. A potential larger-scale study could also assess the impacts of letter grades on high school teachers across several states. The larger samples associated with these proposals would provide more conclusive evidence of whether high school teachers are disproportionately impacted by the letter grade labels. These future studies would also make a substantial contribution to scholarship given that most of the available literature on letter grades in school accountability cited in Chapter 2 focused on elementary and middle schools.

Another important area for study involves extending the impact of the School Performance Grades to explore subsequent student performance. The ultimate goal of school accountability systems is to improve student learning. As outlined in Chapter 2, studies on A-F grading in Florida and New York City showed significant subsequent impacts on elementary and middle school academic achievement on state tests, particularly for schools at the D/F threshold (Figlio & Rouse, 2005; Chiang, 2009; Rockoff & Turner, 2010; Winters & Cowen, 2012). The data available in North Carolina would allow for researchers to replicate study designs from these other areas to see if the letter grades had a similar impact on student achievement in North Carolina.

## **Explanations for and Mechanisms of Teacher Responses**

The pass vs. fail analysis for all school samples and the individual letter grade analysis for elementary and middle schools showed no detectable impacts of the School Performance Grade label a school received on teachers' perceptions or decision to remain teaching in their schools. Thus, no consistent evidence existed the grade a school received had a *universal* impact on teachers across North Carolina. In the following section, I will explore potential explanations for the null findings, connecting them to the literature review and proposing additional research that could provide further explanation of the mechanisms that connect the accountability labels to teacher perceptions. A later section will explore potential differences that account for significant findings on two outcomes for the high school sample.

**Mechanisms to explain null findings.** The potential mechanisms that explain the null findings fall into two main categories: 1) the labels objectively had no direct or indirect impact on teachers for the measured constructs or 2) there may have been an impact, but it was not detected by the research design or measures.

*Information gain from school grades is different for teachers.* Under the first category, one potential explanation relates to information gain. The information gain achieved by releasing a school report card label is likely much higher for people outside of education than for educators working with their schools each day. The information provided was likely not sufficient to directly change teacher perspectives, as the complexity of school environments and factors that contribute to perceptions of environments were not substantially impacted by a report card label in the way that they could impact people without regular touchpoints with their local schools. Prior literature demonstrated the impact of labels on the public (Figlio & Kenny, 2009) and parents (Charbonneau & Van Ryzin, 2012), but not necessarily on people working within

schools with access to additional information about their environments. A lack of direct impact also confirms the findings of Peterson, Henderson, and West (2014) that teachers do not respond strongly to information about schools coming from external sources.

Future research could help to explain how the letter grades impact the opinions of parents and the public on the quality of their schools and their support for teachers. North Carolina does not collect surveys of parental perceptions for educational accountability in the same way they do for teachers. A study measuring the impact on parents or the greater community would require either collecting original data within North Carolina or conducting the analysis on a different state that collects parent surveys as part of their annual school report cards.

*Previous accountability measures absorbed potential impact of grades.* Previous events and shocks in the accountability system, including the introduction of high-stakes testing in NC in the 1990s, AYP designation under NCLB, various re-norming of tests during the decade of the 2000s, and the introduction of more rigorous standards and tests associated with Common Core could have absorbed much of the impact of the NC accountability policy on teacher perceptions. Prior studies in North Carolina demonstrated lower morale and higher stress levels in the initial years of high-stakes accountability (Jones et al., 1999) and unfavorable views of NCLB-related requirements, sanctions, and labels by school leaders (Lyons & Algozzine, 2006).

The letter grades could have made a more discernable impact in the absence of a prior high-stakes accountability system, such as findings in Florida and New York City. This is an important implication for a state or city implementing a high-stakes accountability system for the first time. However, within the United States, NCLB and subsequent federal policy has introduced nearly universal high-stakes accountability throughout the country. Thus, changing a

policy to be letter grade-based rather than another performance indicator format, may be an inconsequential decision when it comes to general impacts on teachers.

A study examining the accountability shock related to the introduction of Common Core could help to explain these results. With the implementation of more rigorous standards and tests, proficiency rates on state tests fell from a median school-level performance composite of 79% in 2011-12 (NC DPI, 2012b) to 42% in 2012-13 (NC DPI, 2014). In the period included in this study on letter grades, the testing requirements for teachers and students did not significantly change; any impact based on testing could have taken place during the shift to Common Core. A thoughtful design to isolate schools experiencing extreme shifts in proficiency rates and its impact on teachers may reveal a significant impact to teacher perceptions prior to the grades, dampening additional influence from the introduction of School Performance Grades.

*More emphasis on stigma than sanctions.* The A-F grades assigned to schools also placed a much heavier emphasis on stigma than sanctions. Connecting to prior literature, the threat of sanctions appears to be a much stronger policy mechanism than the “reform by shame” (Murillo & Flores, 2002) of placing a negative label on a school. Figlio and Rouse (2005) offered evidence from Florida that placement on a critically low-performing list with stigma but no sanctions created changes within schools like policies that coupled placement on the list with sanctions. Bowen & Trivitt (2014) also found that the removal of the threat of vouchers for students in failing schools in 2006 did not affect achievement gains seen in the Florida accountability model. Alternatively, Saw et al. (2017) provided evidence from Michigan that a stigmatizing label alone without sanctions did not impact changes in schools or subsequent student achievement.

As North Carolina continues to build sanctions into the accountability model for a subset of failing schools, there may be additional responses to conditions attached with the labels. When such labels are coupled with sanctions, or consequences of underperformance, schools may respond in other ways that impact work climate and student achievement. One recent example of sanctions occurred with public responses to the threat of the takeover of low-performing elementary schools under a recently-formed Innovative School District (ISD). The legislation requires the ISD to oversee transferring five of the lowest-performing elementary schools (all of whom received grades of F) to a charter organization (NC Innovative School District, 2018).

Even *threats* of meaningful sanctions could strengthen the degree of response to the grades. For example, in Fall 2017, two schools from Durham Public Schools were placed on the potential takeover list, sparking a large outcry from the community in support keeping their schools under local control (Philip, 2017). Future research on public and educator response to the threat of takeover as part of the ISD would provide further understanding of the impact of the accountability system related to the School Performance Grades.

**Additional impact of the grades went undetected.** A second category of potential explanations concedes that there may have been an impact on teachers, but it was not detected for all samples. Potential impacts of these labels that influenced the theory of action came from experimental evidence (Jacobsen, Saultz, & Snyder, 2014; Ladd & Linderholm, 2008). Researchers may be more likely to observe an effect in a controlled environment, particularly if measurement occurs quickly after a stimulus. However, the sparse significant results may indicate that the effect is harder to detect through a natural experiment like the one used for this study on NC School Performance Grades.

*Contamination from multiple signals.* A potential mechanism driving undetected null findings could include contamination of two different grades assigned to schools in 2013-14 and 2014-15 before the last survey. I ran a preliminary analysis for this alternative explanation by re-running the pass/fail analysis excluding the sample to schools whose grades did not change. The exclusion of these schools, representing approximately 28% of the sample, did not yield any additional significant results. A future study could explore the removal of these schools from each of the individual grade-level thresholds.

Schools also received categorical labels related to their EVAAS value-added growth of “Exceeded,” “Met,” or “Did not Meet.” It is possible that these labels had a higher influence on teachers than the letter grade their school received. Future research could explore the impact of the EVAAS labels by exploiting a similar design to the study of letter grades. Each school receives a numeric composite that determines the value-added label assigned to the school. The impact of the EVAAS label could be assessed through a sharp RD at the Exceeded vs. Met and Met vs. Did not Meet thresholds.

Additionally, a school’s EVAAS value-added status can impact their “low performing” status. A school that receives a D or F for their School Performance Grade can avoid placement on the list if they are deemed to “Exceed” growth with their EVAAS composite. This distinction applied to 60 elementary schools, 34 middle schools, and one high school in the sample. Running the models by excluding schools that received “Exceed” status in the elementary and middle school samples at the pass/fail threshold would isolate the impact of “low performing” status. Based on research in Michigan by Saw et al. (2017), this follow-up study would help to isolate the stigma effect of a failing grade from the “low performing” status with more specific sanctions attached.

**Timing.** Two additional explanations concern the timing of the delivery of the letter grades and when survey responses were collected. One possibility involves a decay of an initial response. The null impacts for the outcomes in the elementary and middle school samples suggest that the letter grade label did not have a pervasive impact on teachers' perceptions several months after the school received the grade. To support this possibility, Figlio and Lucas (2004) found evidence in Florida that an initial impact of the A-F school grades on housing prices diminished following the first year. Although not measuring the impact on teachers, the results raise the possibility that an initial impact of the letter grades could decay over time.

Surveying teachers at a time closer to the receipt of the letter grade would offer a clearer perspective on this issue. A field experiment randomizing the timing of asking teachers about their opinions on their schools at different times following the release of grades could help to assess the validity of this explanation. The survey could also include questions about which areas most contributed to their evaluation of the environment.

Alternatively, the letter grades may not have been in place long enough to detect an impact. There exists the possibility that a cumulative effect of the accountability system on teachers could take years to appear, particularly for outcomes related to teacher turnover. Replicating the study with recently-released data from the 2018 NC Teacher Working Conditions survey and the 2016-17 School Performance Grades would help to assess the validity of this argument.

**Targeted questions.** The NC Teacher Working Conditions survey or the specific constructs chosen for this study also may not be sensitive to understand ways in which the A-F grade impacted school environments. More targeted questions related to impacts of the letter grade like those posed to principals by Smith and Imig (2017) are needed to fully understand the

phenomenon of teacher reactions to the labels. A potential explanatory mixed methods study would help to measure this phenomenon. The quantitative data from the School Performance Grade and Teacher Working Conditions survey could be used to identify a sample of teachers for interviews or targeted survey questions related to their perceptions of the letter grade their school received. Additionally, selecting a sample of teachers who have recently left their schools would help to better understand if accountability labels played a role in their decision to leave.

### **Mechanisms to Explain Significant Findings for the High School C/D Threshold**

Analysis of the high school sample revealed evidence that the receiving a letter grade of D as opposed to C impacted teachers' perceptions of support and the accuracy of state assessments to measure students' understanding of standards. Significant findings only for the high school sample required some reflection regarding how these schools and teachers may be different from those in the elementary and middle school samples. One relevant difference between the environments involves the different set of metrics that determine high school grades as opposed to elementary and middle school grades. In addition, high school teachers serve an older group of students who may be more aware of their school's accountability status and the quality of their education. Although speculative, future research could explore any potential impact of student awareness of accountability measures and labels on teachers. Teachers at the high school level may also feel more visibility in their communities than teachers in elementary and middle schools. Particularly in rural counties, districts may have several regional elementary and middle schools but only one or two larger comprehensive high schools.

The literature offers some additional evidence regarding how these groups of teachers are different. A study by Pearson and Moomaw (2005) found no differences in perceptions of autonomy across teaching levels (elementary, middle, and high school). This finding is

consistent with the null findings for all samples and analyses for the outcome regarding perceived autonomy. The concept of parental trust, however, is related to teachers' feelings of perceived support and could account for differences between the samples. Adams and Christensen (2000) noted higher levels of parental trust in elementary school teachers than high school teachers, a difference that may also be present in the study of NC School Performance Grades.

In another exploration of differences between teachers at different levels, Ladd (2011) used data from the 2006 NC Teacher Working Conditions survey to explore structural differences in responses for elementary, middle, and high school teachers and their connection to teacher turnover. One finding concerned the influence of school leadership on teacher working conditions; leadership plays a more important role in turnover decisions in high schools than in elementary or middle schools. Her factor analysis to find underlying constructs in responses also found a different set of working conditions factors for high school teachers than for elementary and middle school teachers. The systematic difference in responses by level of teacher suggest that either high school teachers respond to the survey in different way or have a fundamental difference in the way they perceive their working environments. Regardless, the study provides further evidence that the results could systematically differ depending on the level of school with which a teacher is associated.

### **Recommendations and Policy Implications**

Based on this research, it is important to apply the findings to offer guidance for future policy decisions related to A-F grades on school accountability report cards. As outlined in the previous sections, more study will help to better assess any potential impacts through more direct assessment of the policy. The lack of clear, universal impacts on teachers' perceptions does not

warrant removal of the School Performance Grade as part of the public school report cards in North Carolina. The detected impacts on high school teachers, however, provide some evidence that the School Performance Grade labels are having an adverse impact on feelings of support and belief in the accuracy of state assessments.

This study explored a narrow portion of the School Performance Grade policy – the direct impact on teachers. Prior research has demonstrated that letter grading policies, particularly the receipt of an F grade, can have positive impacts on student achievement in future years (Figlio & Rouse, 2005; Chiang, 2009; Rockoff & Turner, 2010; Winters & Cowen, 2012). The policy question related to student achievement was beyond the scope of this study. However, better understanding the impact on student achievement is crucial to evaluating whether A-F school report cards represent sound policy. If the letter grades positively impact student achievement with minimal adverse impacts on teachers, the performance information format change may be an effective accountability policy.

Although a relatively small impact on only one of three school groups, however, diminished feelings of support could cause long-term damage to the teaching pipeline in North Carolina, causing teachers to leave the profession or choosing a career path outside of teaching. Given the high financial and performance costs associated with teacher turnover (Watlington et al., 2010), policymakers need to continue to ensure accountability systems not only aid student performance but also incentivize teachers to remain growing professionally and teaching in their schools. Potential short-term increases in student achievement measured by standardized tests should not come at the expense of diminishing feelings of professional support.

**Suggestions for future school report cards.** As established in prior literature, letter grades lower the cost of information about public institutions by offering easy-to-understand

information about the performance of taxpayer-funded institutions (Coe & Brunet, 2006).

Multiple measures on report cards, summarized by letter grades, may serve as a good compromise between detractors and proponents of A-F policies. Instead of a single letter grade for each school, offering the public several summary letter grades on a variety of indicators such as test scores, value-added growth, college readiness, career readiness, attendance, and forms of educational attainment would provide a fair examination of the complexity of school performance with an easy-to-understand format for judgments and comparison of public institutions. At minimum, publishing separate grades for achievement and growth for each school would improve this current limitation of the school report cards to accurately convey information.

Policymakers should also exercise more care to ensure that letter grades better allow for consumers to make fair apples-to-apples comparisons. One area for improvement would address norming grades across elementary, middle, and high schools. Each of these school samples has different measures that contribute to the score that determines the grade. This is particularly true for high schools with measures like graduation rate, Math III participation, and the WorkKeys assessment that have much higher means than proficiency on end-of-grade and end-of-course tests. The distribution of letter grades to high schools vs. middle schools does not mean that high school performance is superior to that of middle schools. Based on calculations I made in the analytic file, a grade of C for middle schools represents the 40<sup>th</sup> – 82<sup>nd</sup> percentile. By contrast, a grade of C for high schools shows much lower relative performance in the 11<sup>th</sup> – 63<sup>rd</sup> percentile. The different letter grade distribution is purely symptomatic of the different measures used in the calculations. Thus, if community members or parents form judgments of their local schools

based on the letter grades alone, the measures represent overly simplistic information that masks the underlying performance meanings based on school type.

Another area for improvement involves greater control and transparency for factors outside a school's control. Established by Pierson et al. in a 2015 policy report, the School Performance Grades strongly correlate with poverty. Adding an indicator for schools relative to others with similar levels of economically disadvantaged students would allow for consumers of the school report card to gain a fuller picture of the performance of a school, controlling for external factors to the learning environment. EVAAS value-added growth offers one technique for controlling for external factors, but the complexity of the algorithm makes public interpretability difficult. An additional measure to compare a school's performance with schools with similar poverty levels would provide a more transparent view of the impact of the school environment on student achievement.

The impact of accountability measures and labels on schools and educators is an important area for continued study as, through NCLB and ESSA, performance measurement and public reporting through school report cards remains a relevant issue. Academic measures should help to identify areas for improvement that schools can use for creating better learning opportunities for the students they serve. Whether A-F systems or other means for conveying performance, the information in school report cards should be accurate and should support better school systems. Further understanding of how performance measurement impacts educators is an important consideration in future accountability system design.

## REFERENCES

- Adams, C. M., Dollarhide, E., Forsyth, P. B., Gaetane, J. M., Garland, P., Miskell, R., Mwavita, M. (2013). *An examination of the Oklahoma State Department of Education's A-F report card*. Retrieved from <http://okea.org/assets/files/A-F Study.pdf>
- Arkansas News. (2015, April 16). Arkansas schools get first A-F report cards. *Arkansas News*. Retrieved from <http://www.arkansasnews.com/news/arkansas/arkansas-schools-get-first-f-report-cards>
- Bowen, D. H., & Trivitt, J. R. (2014). Stigma without sanctions: The (lack of) impact of private school vouchers on student achievement. *Education Policy Analysis Archives*, 22(87), 1–37.
- Burkhauser, S. (2017). How much do school principals matter when it comes to teacher working conditions? *Educational Evaluation and Policy Analysis*, 39(1), 126–145.
- Charbonneau, E., & Van Ryzin, G. G. (2012). Performance measures and parental satisfaction with New York City schools. *The American Review of Public Administration*, 42(1), 54–65.
- Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics*, 93(9–10), 1045–1057.
- Chingos, M. M., Henderson, M., & West, M. R. (2012). Citizen perceptions of government service quality: Evidence from public schools. *Quarterly Journal of Political Science*, 7(4), 411–445.
- Clotfelter, C. T., Ladd, H. F., Vigdor, J. L., & Diaz, R. A. (2004). Do school accountability systems make it more difficult for low-performing schools to attract and retain high-quality teachers? *Journal of Policy Analysis and Management*, 23(2), 251–271.
- Coe, C. K., & Brunet, J. R. (2006). Organizational report cards: Significant impact or much ado about nothing? *Public Administration Review*, 66(1), 90–100.

- Crain, T. P. (2016a, September 14). Alabama's A-F school grading system is almost ready. *Alabama School Connection*. Retrieved from <http://alabamaschoolconnection.org/2016/09/14/alabamas-a-f-school-grading-system-is-almost-ready/>
- Crain, T. P. (2016b, November 10). No letter grades for Alabama schools this year. *AL.com*. Retrieved from [http://www.al.com/news/index.ssf/2016/11/no\\_letter\\_grades\\_for\\_alabama\\_s.html](http://www.al.com/news/index.ssf/2016/11/no_letter_grades_for_alabama_s.html)
- Crain, T. P. (2018, February 1). Alabama public school grades released: Find yours here. *AL.com*. Retrieved from [https://www.al.com/news/index.ssf/2018/02/alabamas\\_k-12\\_public\\_school\\_gr.html](https://www.al.com/news/index.ssf/2018/02/alabamas_k-12_public_school_gr.html)
- Dimmery, D. (2016). Package "rdd": Regression discontinuity estimation. Retrieved from <https://cran.r-project.org/web/packages/rdd/rdd.pdf>
- Dizon-Ross, R. (2014). How do school accountability reforms affect teachers? Evidence from New York City. Retrieved from [http://scholar.harvard.edu/files/rdr/files/accountability\\_and\\_teachers\\_2014feb4.pdf](http://scholar.harvard.edu/files/rdr/files/accountability_and_teachers_2014feb4.pdf)
- Elliott, S. (2013, December 22). The basics of A-F grading in Indiana: Changes and controversy. *Chalkbeat*. Retrieved from <https://www.chalkbeat.org/posts/in/2013/12/22/the-basics-of-a-to-f-grading-in-indiana/>
- ExcelinEd. (2015, September 30). Utah: Raising school grades and expectations. Retrieved from <http://www.excelined.org/2015/09/30/utah-raising-school-grades-and-expectations/>
- Favero, N., & Meier, K. J. (2013). Evaluating urban public schools: Parents, teachers, and state assessments. *Public Administration Review*, 73(3), 401–412.
- Feng, L., Figlio, D. N., & Sass, T. (2010). *School accountability and teacher mobility* (NBER

- Working Paper No. 16070). Retrieved from [www.nber.org/papers/w16070](http://www.nber.org/papers/w16070)
- Figlio, D., & Loeb, S. (2011). School accountability. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the economics of education* (Vol. 3, pp. 384–421). Amsterdam: Elsevier.
- Figlio, D. N., & Kenny, L. W. (2009). Public sector performance measurement and stakeholder support. *Journal of Public Economics*, 93(9–10), 1069–1077.
- Figlio, D. N., & Lucas, M. E. (2004). What's in a grade? School report cards and the housing market. *American Economic Review*, 94(3), 591–604.
- Figlio, D. N., & Rouse, C. E. (2005). *Do accountability and voucher threats improve low-performing schools?* (NBER Working Paper No. 11597). Cambridge, MA: National Bureau of Economic Research. Retrieved from <http://www.nber.org/papers/w11597>
- Finnigan, K. S., & Gross, B. (2007). Do accountability policy sanctions influence teacher motivation? Lessons from Chicago's low-performing schools. *American Educational Research Journal*, 44(3), 594–630.
- Friesen, J., Javdani, M., Smith, J., & Woodcock, S. (2012). How do school “report cards” affect school choice decisions? *Canadian Journal of Economics*, 45(2), 784–807.
- Gershenson, S. (2016). Performance standards and employee effort: Evidence from teacher absences. *Journal of Policy Analysis and Management*, 35(3), 615–638.
- Goldhaber, D. D., & Hannaway, J. (2004). Accountability with a kicker: Observations on the Florida A+ accountability plan. *Phi Delta Kappan*, 85(8), 598.
- Grissom, J. A., Nicholson-Crotty, S., & Harrington, J. R. (2014). Estimating the effects of No Child Left Behind on teachers' work environments and job attitudes. *Educational Evaluation and Policy Analysis*, 36(4), 417–436.

- Hastings, J. S., & Weinstein, J. M. (2008). Information, school choice, and academic achievement: Evidence from two experiments. *Quarterly Journal of Economics*, *123*(4), 1373–1414.
- Heissel, J. A., & Ladd, H. F. (2016). *School turnaround in North Carolina: A regression discontinuity analysis* (CALDER Working Paper No. 156). Washington, DC: American Institutes for Research. Retrieved from <https://caldercenter.org/sites/default/files/WP%20156.pdf>
- Howe, K. R., & Murray, K. (2015). *Why school report cards merit a failing grade*. Boulder, CO: National Education Policy Center. Retrieved from <http://nepc.colorado.edu/publication/why-school-report-cards-fail>
- Hu, L. T. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1-55.
- Imbens, G. W., & Kalyanaraman, K. (2009). *Optimal bandwidth choice for the regression discontinuity estimator* (NBER Working Paper No. 14726). Cambridge, MA: National Bureau of Economic Research. Retrieved from <http://www.nber.org/papers/w14726.pdf>
- Jacobsen, R., Saultz, A., & Snyder, J. W. (2013). When accountability strategies collide: Do policy changes That raise accountability standards also erode public satisfaction? *Educational Policy*, *27*(2), 360–389.
- Jacobsen, R., Snyder, J. W., & Saultz, A. (2014). Informing or shaping public opinion? The influence of school accountability data format on public perceptions of school quality. *American Journal of Education*, *121*(1), 1–27.
- 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.
- Jones, M. G., Jones, B. D., Hardin, B., Chapman, L., Yarbrough, T., & Davis, M. (1999). The impact of high-stakes testing on teachers and students in North Carolina. *The Phi Delta Kappan*, 81(3), 199–203.
- King-Sears, M. E., & Baker, P. H. (2014). Comparison of teacher motivation for mathematics and special educators in middle schools that have and have not achieved AYP. *ISRN Education*, 2014, 1–12.
- Ladd, H. F., & Glennie, E. (2001). A replication of Jay Greene's voucher effect study using North Carolina data. In M. Carnoy (Ed.), *Do School vouchers Improve Student Performance?* Washington, D.C.: Economic Policy Institute.
- Ladd, H. F. (2011). Teachers' perceptions of their working conditions: How predictive of planned and actual teacher movement? *Educational Evaluation and Policy Analysis*, 33(2), 235-261.
- Ladd, J. A., & Linderholm, T. (2008). A consequence of school grade labels: Preservice teachers' interpretations and recall of children's classroom behavior. *Social Psychology of Education*, 11(3), 229–241.
- Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2), 281–355.
- Leithwood, K., & Mcadie, P. (2010). Teacher working conditions that matter. *Education Canada*, 47(2), 42–45.
- Loeb, S., Darling-Hammond, L., & Luczak, J. (2005). How teaching conditions predict teacher turnover in California schools. *Peabody Journal of Education*, 80(3), 44–70.
- Louisiana Department of Education. (2013). School and district report cards. Retrieved from

<https://www.louisianabelieves.com/data/reportcards/>

- Lyons, J. E., & Algozzine, B. (2006). Perceptions of the impact of accountability on the role of principals. *Education Policy Analysis Archives*, 14(16), 1–16. Retrieved from <http://files.eric.ed.gov/fulltext/EJ806035.pdf>
- Maddock, A. (2009). *North Carolina teacher working conditions: The intersection of policy and practice*. Santa Cruz, CA: New Teacher Center. Retrieved from [http://www.jntp.org/sites/default/files/ntc/main/pdfs/NC\\_TWC\\_Policy\\_Practice.pdf](http://www.jntp.org/sites/default/files/ntc/main/pdfs/NC_TWC_Policy_Practice.pdf)
- Maine Department of Education. (2015). Report cards for Maine schools, transparency for Maine people: The Maine School Performance Grading System. Retrieved from <http://maine.gov/doe/schoolreportcards/>
- Martriano, M. J., & Green, M. I. (2016). West Virginia A-F school accountability system: West Virginia's school report cards. Retrieved from [https://static.k12.wv.us/a-f/a-f\\_aboutwithgrades.pdf](https://static.k12.wv.us/a-f/a-f_aboutwithgrades.pdf)
- Mintrop, H., & Trujillo, T. (2005). Corrective action in low performing schools: Lessons for NCLB implementation first-generation accountability systems. *Education Policy Analysis Archives*, 13(48). Retrieved from <http://epaa.asu.edu/epaa/v13n48/>
- Mississippi Center for Public Policy. (2012, September 17). New A-F grades for 2011-2012. Retrieved from <http://www.msppolicy.org/new-a-f-grades-for-2011-2012/>
- Moynihan, D. P. (2008). *Public management and change: Dynamics of performance management: Constructing information and reform*. Washington, D.C.: Georgetown University Press.
- Murillo, E., & Flores, S. (2002). Reform by shame: Managing the stigma of labels in high stakes testing. *Educational Foundations*, 16(2), 93–108.

Murray, J. (2013, September 9). Parsing performance analysis of Ohio's new school report cards.

Retrieved from <https://edexcellence.net/parsing-performance-analysis-of-ohio's-new-school-report-cards>

NC Innovative School District (2018). *What is the NC Innovative School District?* Retrieved

from <http://innovativeschooldistrict.org/about/what-is-innovative-school-district/>

New Mexico Public Education Department. (2015). A-F school grading: Frequently asked

questions. Retrieved from

[https://aae.ped.state.nm.us/SchoolGradingLinks/1516/TECHNICAL ASSISTANCE FOR EDUCATORS/School Grading FAQs.pdf](https://aae.ped.state.nm.us/SchoolGradingLinks/1516/TECHNICAL_ASSISTANCE_FOR_EDUCATORS/School_Grading_FAQs.pdf)

New Teacher Center (2018). *2016 North Carolina Teacher Working Conditions survey: Student*

*achievement and teacher retention analyses*. Retrieved from

[https://ncteachingconditions.org/uploads/File/NC16\\_report\\_final.pdf](https://ncteachingconditions.org/uploads/File/NC16_report_final.pdf)

North Carolina Department of Public Instruction (2012). *The ABCs accountability model:*

*Performance of all schools 2011-2012*. Retrieved from

<http://abcs.ncpublicschools.org/abcs/abcVol1List.jsp?pYear=2011-2012&pList=5&pListVal=2&GO=GO>

North Carolina Department of Public Instruction (2012, August 2). *Evolution of the ABCs*.

Retrieved from <http://www.ncpublicschools.org/docs/accountability/reporting/abc/2011-12/abcevolution.pdf>

North Carolina Department of Public Instruction (2014). *2012-13 state, district, and school level*

*summary data*. Retrieved from

<http://www.ncpublicschools.org/accountability/reporting/ncaccmodel.zip>

- North Carolina Department of Public Instruction (2015, February 5). *2013-14 school performance grades (A-F) for North Carolina public schools*. Retrieved from <http://www.ncpublicschools.org/docs/accountability/reporting/spgexecsumm15.pdf>
- North Carolina Department of Public Instruction (2016, August 29). *North Carolina data release technical notes: 2015-16 school year*. Retrieved from <http://www.ncpublicschools.org/docs/accountability/reporting/datarlstchnts16.pdf>
- North Carolina Department of Public Instruction (2018). *Low-performing school and district plans*. Retrieved from <http://www.dpi.state.nc.us/schooltransformation/low-performing/>
- Oklahoma Department of Education. (2015). A to F report card calculation guide. Retrieved from <http://sde.ok.gov/sde/sites/ok.gov.sde/files/documents/files/AtoFReportCardGuide.pdf>
- Olsen, A. L. (2013). Leftmost-digit-bias in an enumerated public sector? An experiment on citizens' judgment of performance information. *Judgment and Decision Making*, 8(3), 365–371.
- Palmer, J. (2016). Arizona state board of education adopts new A-F school accountability plan. Retrieved from <http://www.helios.org/blog/arizona-state-board-of-education-adopts-new-a-f-school-accountability-system>
- Peterson, P. E., Henderson, M. B., & West, M. R. (2014). *Teachers versus the public: What Americans think about their schools and how to fix them*. Washington, DC: Brookings Institution Press.
- Philip, L. (2017, October 2). Durham parents feel 'taken over' by school takeover process [Radio broadcast]. In F. Stasio (Host), *The State of Things*. Durham, NC: WUNC. Retrieved from <http://wunc.org/post/durham-parents-feel-taken-over-school-takeover-process>
- Pierson, J. B., Maugeri, J., Reitano, V., & Xing, Q. W. (2015). *Grading school performance*

- grades: A preliminary analysis of the existing system and recommendations to improve transparency and support*. Raleigh, NC: NC Department of Public Instruction. Retrieved from <http://www.ncpublicschools.org/docs/intern-research/reports/gradingspg2015.pdf>
- Reed, C. J., McDonough, S., Ross, M., & Robichaux, R. (2001). *Principals' perceptions of the impact of high stakes testing on empowerment*. Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA. Retrieved from <http://files.eric.ed.gov/fulltext/ED459538.pdf>
- Revelle, W. (2017). Package “psych”: Procedures for psychological, psychometric, and personality research. Retrieved from <https://cran.r-project.org/web/packages/psych/psych.pdf>
- Rockoff, J., & Turner, L. J. (2010). Short-run impacts of accountability on school quality. *American Economic Journal: Economic Policy*, 2(4), 119–147.
- Rosseel, Y., Byrnes, D., Vanbrabant, L., Savalei, V., Merkle, E., & Hallquist, M. (2017). Package “lavaan”: Latent variable analysis. Retrieved from <https://cran.r-project.org/web/packages/lavaan/lavaan.pdf>
- Rouse, C. E., Hannaway, J., Goldhaber, D., & Figlio, D. (2013). Feeling the florida heat? How low-performing schools respond to voucher and accountability pressure. *American Economic Journal: Economic Policy*, 5(2), 251–281.
- Sabin, J. T. (2015). Teacher morale, student engagement, and student achievement growth in reading: A correlational study. *Journal of Organizational & Educational Leadership*, 1(1). Retrieved from <http://digitalcommons.gardner-webb.edu/joel/vol1/iss1/5/>
- Saw, G., Schneider, B., Frank, K., Chen, I.-C., Keesler, V., & Martineau, J. (2017). The impact of being labeled as a persistently lowest achieving school: Regression discontinuity

- evidence on school sanctions. *American Journal of Education*, 123, 585–613.
- Smith, R., & Imig, S. R. (2017). The fallacy of school grades: Exploring the myth that public shaming leads to school improvement. In C. Meyers & M. Darwin (Eds.), *Enduring myths that inhibit school turnaround* (pp. 297–317). Charlotte, NC: Information Age Publishing.
- Sun, M., Saultz, A., & Ye, Y. (2016). Federal policy and the teacher labor market: Exploring the effects of NCLB school accountability on teacher turnover. *School Effectiveness and School Improvement*, 28(1), 102–122.
- Tanner, J. (2016). *The A-F accountability mistake*. The Texas Accountability Series. Austin, TX: The Texas Association of School Administrators. Retrieved from <https://www.tasanet.org/cms/lib07/TX01923126/Centricity/Domain/393/A-F-Essay.pdf>
- The Georgia Governor's Office of Student Achievement. (2017). School-level data. Retrieved from <https://schoolgrades.georgia.gov/dataset/school-level-data>
- U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse (2013). *What works clearinghouse v.3.0 standards handbook*. Retrieved from [https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc\\_procedures\\_v3\\_0\\_standards\\_handbook.pdf](https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v3_0_standards_handbook.pdf)
- United States Department of Education (2017). *Every Student Succeeds Act state and local report cards non-regulatory guidance*. Retrieved from <https://www2.ed.gov/policy/elsec/leg/essa/essastatereportcard.pdf>
- Watlington, E., Shockley, R., Guglielmino, P., & Felsher, R. (2010). The high cost of leaving: An analysis of the cost of teacher turnover. *Journal of Education Finance*, 36(1), 22–37.
- Winters, M. A. (2016). *Grading schools promotes accountability and improvement: Evidence*

*from New York City, 2013-15*. New York, NY: Manhattan Institute. Retrieved from <https://www.manhattan-institute.org/html/grading-schools-promotes-accountability-and-improvement-evidence-nyc-2013-15-8912.html>

Winters, M. A., & Cowen, J. M. (2012). Grading New York: Accountability and Student Proficiency in America's Largest School District. *Educational Evaluation and Policy Analysis*, 34(3), 313–327.

## APPENDICES

## Appendix A

I used assumptions of  $\alpha = .05$ , a two-tailed test, and power of  $(1 - \beta) = .80$  using the PowerUp macro-enabled spreadsheet (Dong & Maynard, 2013) as the basis for all calculations. Given the usage of teacher-level responses, I used the tab for “Model 5.3: MDES Calculator for Two-Level Regression Discontinuity Designs – Treatment at Level 2” (Dong & Maynard, 2013) to provide the appropriate formulas for the power calculations. The tab combines parameters for a standard two-level power analysis with the RD Design Factor outlined by Schochet (2008). This design matched the data because I explored individual teacher-level responses (Level 1) with the treatment of School Performance Grade assignment to schools (Level 2).

The parameters used for power analysis required many different calculations for each sample, outcome, and bandwidth. Conducting all power analysis within the PowerUp spreadsheet was challenging to keep track of all the parameters. To address the issue, I wrote functions in R to conduct the RD power analysis for a variety of parameters associated with the different analyses. I built the functions from equations in the PowerUp spreadsheet and from Schochet’s 2008 paper on the RD Design Effect.

The bandwidth around each letter grade cutoff represented a different  $n$ -size for each grade-level sample. I calculated the prospective sample sizes and corresponding power analysis in an Excel spreadsheet. Table A.1 lists the  $n$ -sizes associated with each cutoff for four proposed bandwidths for each sample. I considered a maximum bandwidth of 15 points on either side of the cutoff because each letter grade only represents a range of 15 points, as shown in Table 1.2.

Table A.1  
*Sample Size of Schools at Each Cutoff for Each Sample by Four Selected RD Bandwidths*

		5 Point Bandwidth	7 Point Bandwidth	10 Point Bandwidth	15 Point Bandwidth
Elementary	A/B Cutoff	60	95	156	375
	B/C Cutoff	314	432	585	829
	C/D Cutoff	319	416	570	828
	D/F Cutoff	108	171	227	362
Middle	A/B Cutoff	14	23	42	198
	B/C Cutoff	91	131	188	267
	C/D Cutoff	128	167	235	339
	D/F Cutoff	50	78	108	174
High	A/B Cutoff	24	40	69	123
	B/C Cutoff	117	166	219	281
	C/D Cutoff	68	93	137	205
	D/F Cutoff	4	6	12	37

PowerUp does not calculate the value of  $\rho_{TS}$ , the correlation between the treatment indicator and the score used for assignment to treatment. Knowing this value allows the calculation of the RD Design Effect multiplier for the MDES. For the proposed study, this parameter is the correlation between the discrete value of the School Performance Grade at each cutoff (i.e., the treatment) and the value of the School Performance Grade score (i.e., the running variable). I used the formula provided by Schochet (2008) for a design that fits scenario three, an aggregated design in which “schools are the unit of assignment and no random classroom effects” (p. 5). The distribution of the running variable, illustrated in Figure A.1, also neatly fits the description of a truncated normal distribution because the School Performance Grade score is normally distributed and the cutoff scores for each letter grade threshold represent different segments of the distribution. I used the formula for the “truncated normal distribution” (p. 12) to calculate the RD Design Effect independently for each potential bandwidth and each grade cutoff. Table A.2 contains the RD Design Effect

parameters used in the calculations. The factor decreases as the bandwidth increases, increasing the statistical power of the design.

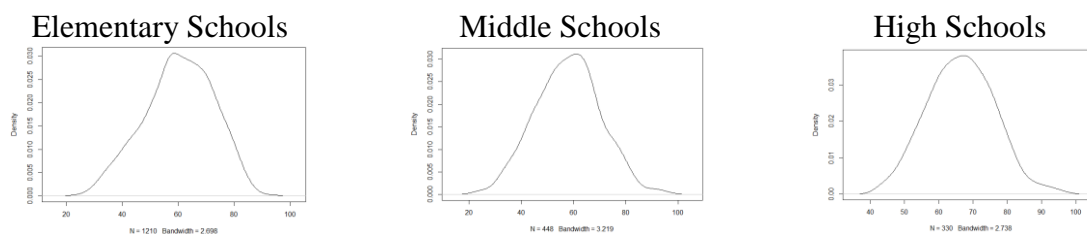


Figure A.1. Distribution of the SPG score, the running variable for the RD.

Table A.2

*RD Design Effect Parameter Estimates at each Cutoff for Four Selected Bandwidths for Each Sample*

		5 Point Bandwidth	7 Point Bandwidth	10 Point Bandwidth	15 Point Bandwidth
Elementary	A/B Cutoff	1.35	1.18	1.12	1.47
	B/C Cutoff	3.69	3.38	2.87	2.36
	C/D Cutoff	2.78	2.41	2.73	2.61
	D/F Cutoff	2.54	1.82	1.76	1.44
Middle	A/B Cutoff	1.15	1.37	1.21	1.09
	B/C Cutoff	2.67	2.41	2.36	2.11
	C/D Cutoff	3.89	3.49	2.49	2.47
	D/F Cutoff	2.27	2.19	1.50	1.32
High	A/B Cutoff	1.55	1.38	1.24	1.13
	B/C Cutoff	3.45	3.25	3.02	2.67
	C/D Cutoff	3.17	2.57	1.99	1.66
	D/F Cutoff	N/A	N/A	N/A	N/A

I used the 2016 NC Teacher Working Conditions data set to calculate additional data-specific parameters for the power analysis. The necessary parameters for the power analysis included the intraclass correlation coefficient (*ICC*), proportion of schools receiving treatment (*P*), the  $R^2$  value for school-level covariates, the number of school-level covariates ( $g^*$ ), the average number of teachers per school ( $n$ ), and the number of schools ( $J$ ). Table A.3 contains MDES calculations by grade cutoff for four different bandwidths using values from the 2014-15 School Performance Grade data file and the 2016 administration of the NC Teacher Working Conditions Survey.

Table A.3  
*Minimum Detectable Effect Size (MDES) for the Pass/Fail Analysis by Teacher Construct*

	Elementary Schools	Middle Schools	High Schools
Support	0.10	0.14	0.17
Autonomy	0.11	0.16	0.19
Accuracy of State Assessments	0.08	0.10	0.13
Good Place to Work and Learn	0.10	0.14	0.17
Intend to Stay Teaching at Same School	0.08	0.11	0.12
Intend to Leave Education Entirely	0.06	0.07	0.10

All MDES values for the pass/fail analysis were less than  $.20 SD$ . For context, each outcome variable is measured on a four-point Likert scale from Strongly Disagree to Strongly Agree. With responses coded 1-4, prior year data indicated a teacher-level standard deviation of approximately 0.75 for any given item. Thus, a  $.20 SD$  impact would indicate that, on average, about 15% of staff members rated an item one scale point lower at the cutoff between “failing” and “non-failing” schools. If the letter grades impacted teacher perceptions, such a difference appeared reasonable to detect.

Tables A.4, A.5, and A.6 contain the MDES values for the elementary school samples at each letter grade cutoff and selected bandwidths. In evaluating the appropriateness of an RD at each cutoff, a tradeoff existed between restricting the sample to values as close to the cutoff as possible and a large enough sample size to adequately detect effects. A 10-point bandwidth offered acceptable MDES values but did not encroach into values for other grade bands. For the main analysis, I evaluated the impact for a 10-point and 15-point bandwidth as a robustness check for all samples. For potentially significant findings, I ran additional models using a 5-point and 7-point bandwidth, which are detailed in the section on sensitivity analysis in Chapter 4.

Table A.4  
*Minimum Detectable Effect Size (MDES) at each Cutoff for Four Selected Bandwidths by Teacher Perception Construct for the Elementary School Sample*

		5 Point	7 Point	10 Point	15 Point
		Bandwidth	Bandwidth	Bandwidth	Bandwidth
Support	A/B	0.48	0.39	0.33	0.28
	B/C	0.21	<b>0.18</b>	<b>0.15</b>	<b>0.13</b>
	C/D	<b>0.20</b>	<b>0.16</b>	<b>0.14</b>	<b>0.11</b>
	D/F	0.29	0.22	<b>0.17</b>	<b>0.14</b>
Autonomy	A/B	0.57	0.46	0.39	0.33
	B/C	0.25	0.21	<b>0.18</b>	<b>0.15</b>
	C/D	0.23	<b>0.19</b>	<b>0.16</b>	<b>0.13</b>
	D/F	0.34	0.27	<b>0.20</b>	<b>0.16</b>
Accuracy of State Assessments	A/B	0.42	0.34	0.29	0.24
	B/C	<b>0.18</b>	<b>0.15</b>	<b>0.13</b>	<b>0.11</b>
	C/D	<b>0.17</b>	<b>0.14</b>	<b>0.12</b>	<b>0.09</b>
	D/F	0.25	<b>0.20</b>	<b>0.15</b>	<b>0.12</b>
Good Place to Work and Learn	A/B	0.50	0.40	0.34	0.28
	B/C	0.21	<b>0.18</b>	<b>0.16</b>	<b>0.13</b>
	C/D	<b>0.20</b>	<b>0.17</b>	<b>0.14</b>	<b>0.11</b>
	D/F	0.30	0.23	<b>0.17</b>	<b>0.14</b>
Intend to Stay at Same School	A/B	0.25	0.20	0.21	<b>0.10</b>
	B/C	<b>0.15</b>	<b>0.12</b>	<b>0.11</b>	<b>0.09</b>
	C/D	<b>0.14</b>	<b>0.11</b>	<b>0.11</b>	<b>0.10</b>
	D/F	0.25	<b>0.16</b>	<b>0.16</b>	<b>0.13</b>
Intend to Leave Education Entirely	A/B	0.22	<b>0.19</b>	<b>0.20</b>	<b>0.09</b>
	B/C	<b>0.11</b>	<b>0.10</b>	<b>0.09</b>	<b>0.07</b>
	C/D	<b>0.09</b>	<b>0.07</b>	<b>0.08</b>	<b>0.07</b>
	D/F	<b>0.17</b>	<b>0.11</b>	<b>0.12</b>	<b>0.09</b>

Note: Bolded values indicate an MDES of .20 *SD* or less.

Table A.5  
*Minimum Detectable Effect Size (MDES) at each Cutoff for Four Selected Bandwidths by Teacher Perception Construct for the Middle School Sample*

		5 Point Bandwidth	7 Point Bandwidth	10 Point Bandwidth	15 Point Bandwidth
Support	A/B	N/A	0.49	0.31	0.18
	B/C	0.29	0.22	<b>0.18</b>	<b>0.15</b>
	C/D	0.30	0.25	<b>0.17</b>	<b>0.14</b>
	D/F	0.39	0.30	0.21	<b>0.18</b>
Autonomy	A/B	N/A	0.82	0.56	0.30
	B/C	0.39	0.32	0.26	0.21
	C/D	0.36	0.29	0.20	<b>0.17</b>
	D/F	0.46	0.36	0.26	0.21
Accuracy of State Assessments	A/B	N/A	0.53	0.34	<b>0.16</b>
	B/C	0.25	0.20	<b>0.15</b>	<b>0.13</b>
	C/D	0.25	0.20	<b>0.13</b>	<b>0.11</b>
	D/F	0.36	0.28	<b>0.20</b>	<b>0.16</b>
Good Place to Work and Learn	A/B	N/A	0.57	0.38	0.22
	B/C	0.32	0.25	0.21	<b>0.17</b>
	C/D	0.33	0.27	<b>0.18</b>	<b>0.16</b>
	D/F	0.45	0.36	0.26	<b>0.20</b>
Intend to Stay at Same School	A/B	N/A	0.49	0.31	<b>0.15</b>
	B/C	0.24	0.20	<b>0.16</b>	<b>0.13</b>
	C/D	0.26	0.25	<b>0.17</b>	<b>0.12</b>
	D/F	0.37	0.29	0.20	<b>0.16</b>
Intend to Leave Education Entirely	A/B	N/A	0.48	0.30	<b>0.13</b>
	B/C	0.20	<b>0.16</b>	<b>0.12</b>	<b>0.10</b>
	C/D	<b>0.19</b>	0.23	<b>0.16</b>	<b>0.08</b>
	D/F	0.25	<b>0.19</b>	<b>0.12</b>	<b>0.10</b>

Note: Bolded values indicate an MDES of .20 *SD* or less.

Table A.6  
*Minimum Detectable Effect Size (MDES) at each Cutoff for Four Selected Bandwidths by Teacher Perception Construct for the High School Sample*

		5 Point Bandwidth	7 Point Bandwidth	10 Point Bandwidth	15 Point Bandwidth
Support	A/B	0.52	0.37	0.30	0.26
	B/C	0.26	0.20	<b>0.18</b>	<b>0.16</b>
	C/D	0.35	0.27	0.22	<b>0.19</b>
	D/F	N/A	N/A	N/A	N/A
Autonomy	A/B	0.70	0.52	0.42	0.35
	B/C	0.31	0.24	0.21	<b>0.18</b>
	C/D	0.38	0.32	0.26	<b>0.20</b>
	D/F	N/A	N/A	N/A	N/A
Accuracy of State Assessments	A/B	0.46	0.33	0.28	0.24
	B/C	0.22	<b>0.18</b>	<b>0.16</b>	<b>0.14</b>
	C/D	0.30	0.24	<b>0.19</b>	<b>0.16</b>
	D/F	N/A	N/A	N/A	N/A
Good Place to Work and Learn	A/B	0.57	0.41	0.34	0.29
	B/C	0.27	0.21	<b>0.18</b>	<b>0.16</b>
	C/D	0.37	0.29	0.24	<b>0.20</b>
	D/F	N/A	N/A	N/A	N/A
Intend to Stay at Same School	A/B	0.41	0.28	0.24	0.21
	B/C	0.21	<b>0.16</b>	<b>0.15</b>	<b>0.13</b>
	C/D	0.29	0.24	<b>0.20</b>	<b>0.16</b>
	D/F	N/A	N/A	N/A	N/A
Intend to Leave Education Entirely	A/B	0.39	0.27	0.23	0.21
	B/C	<b>0.19</b>	<b>0.15</b>	<b>0.14</b>	<b>0.12</b>
	C/D	0.25	0.21	<b>0.18</b>	<b>0.14</b>
	D/F	N/A	N/A	N/A	N/A

Note: Bolded values indicate an MDES of .20 *SD* or less.