The Impact of NC School Performance Grades on Teacher Perceptions and Intent to Leave their Schools

Eric M. Grebing

North Carolina State University

Abstract

The study used two-level regression discontinuity models to evaluate the impact of the NC School Performance (A-F) Grades on teacher perceptions regarding support, autonomy, state assessments, work climate, and intent to remain teaching in their schools. The timing of the study exploited a natural experiment that isolated the impact of the grade label independent of other changes to the accountability model. 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 elementary and middle school samples. Additional exploration of the B/C, C/D, and D/F cutoffs for elementary schools also showed no significant discontinuities associated with the labels. A follow-up sensitivity analysis excluding schools with different grades between 2013-14 and 2014-15 also revealed no significant impacts. The lack of detected impact attributed to the letter grades could be attributed to a variety of factors that will require additional analysis to isolate.

Background

State accountability systems often attach labels to school performance. These labels, meant to hold schools accountable and improve educational outcomes for students, lead to intended and unintended impacts. Teacher perceptions and choosing to leave their schools are potential consequences of the assigned labels. 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 and immediate professional plans by combining administrative data and responses from the biannual North Carolina Teacher Working Conditions Survey.

North Carolina school accountability. The ABCs plan, an early form of 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). North Carolina assigned performance labels 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 January 2018, 14 states used A-F grades to label schools on annually-released school report cards. Table 1 summarizes the states with current A-F grading policies. Two additional states, Maine and Alabama, formerly used A-F school grades but abandoned the practice within the last five years. 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 in other states through understanding consequences of A-F school performance labels.

TABLE 1 ABOUT HERE

Benefits and drawbacks 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). Particularly for schools receiving “failing” grades in the form of a D or F, the label can negatively impact 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' perceptions of 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 setting 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 and behavioral changes associated with the designation. The introduction of these labels serves as a natural experiment for schools that were not previously identified as failing that now had a D or F label attached to them. Figure 1 describes the timing of events for the proposed study.

FIGURE 1 ABOUT HERE

North Carolina School Performance Grades. Passed in 2013 by the NC General Assembly, G.S. 115C-83.15(b) articulated the components used to calculate School Performance Grades. As of September 2017, the calculation of grades still followed 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 2 contains the conversion values of the School Performance Grade score to a letter grade.

TABLE 2 ABOUT HERE

The state uses different measures to create the composite score depending on whether they describe elementary, middle, or high school grade levels. The grades of elementary and middle schools, the samples for this study, are solely based on test scores for the academic achievement composite but differ depending on the grade levels a school serves. Table 3 details the measures used to determine the grade for each school type. Due to these differences, the study examined impacts separately for elementary and middle school samples.

TABLE 3 ABOUT HERE

NC Teacher Working Conditions Survey. Outcome variables for each research question came from the NC Teacher Working Conditions Survey results. North Carolina surveys all teachers every two years using the NC Teacher Working Conditions Survey (Maddock, 2009). With an 89% response rate from certified staff across the state in 2014 (New Teacher Center, 2014), the survey results offer a biennial snapshot of teacher perceptions for nearly every school in the state. Thus, non-response bias did not pose a significant threat to inference (Groves et al., 2009). 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. Thus, I selected eight conceptually-relevant items from the survey, grouped into two three-item scales and two single-item outcomes.

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 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 exists on the impacts of the School Performance Grade policy in North Carolina (Pierson et al., 2015; Smith & Imig, 2017). 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. However, no research to date aids understanding of how the school-level A-F grades teachers' perceptions of their work environments in North Carolina. 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

Teachers' 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).

Research Questions and Hypotheses

The paper 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) overall school climate?
2) Does the School Performance Grade impact teachers' immediate professional plans?

I hypothesized that the receipt of a D or F grade in 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 in 2014-15 decreased the rate of teachers planning to stay in their schools as indicated by the same survey.

Conceptual Framework

The following conceptual framework guided the investigation, summarized in Figure 2. 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 received a failing grade.

FIGURE 2 ABOUT HERE

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 potentially 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' beliefs that state assessments accurately gauged 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.

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).

Methods

The study design exploits a natural experiment due to the timing of the release of school accountability measures and NC Teacher Working Conditions Survey data, illustrated in Figure 1. In North Carolina school accountability, student performance measures have been tracked and disseminated for nearly two decades (NC DPI, 2012), but the state did not previously assign a summary A-F letter grade to rate the performance of each school. The assignment of North Carolina School Performance Grades therefore represented an exogenous shock because the assignment of grades was outside of a school's control and, outside of improving student performance on tests and high school completion, a school did not have a direct mechanism to manipulate the grade received. 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. 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.

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 (prior to treatment) to improve statistical power.

For each of the two school-type samples and for each of the research questions, I assessed the impact at the cutoff for the full sample between the grade of C and D, representing the difference between a "passing" and "failing" grade. For elementary schools, I conducted an additional analysis to assess the more specific impacts of i) B vs. C, ii) C vs. D, and iii) D vs. F School Performance Grades. The size of the elementary school sample was sufficiently large to detect impacts at each of these cutoffs whereas the middle school sample only allowed for analysis of the "passing" vs. "failing" schools. In addition, very few elementary schools received a grade of A, making the available sample size too small to detect an impact at the A vs. B cutoff. A later section on power analysis describes the process for determining the minimum detectable effect sizes (MDES) for various criteria.

For the first research question, I assessed 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."

Sample selection. NC Teacher Working Conditions data are made publicly available online at the school level only when at least five teachers responded and the response rate is greater than 40% (New Teacher Center, 2014). Thus, I eliminated schools from the sample that

did not meet these same criteria for each year of survey results. I used listwise deletion to handle missing data for all predictor variables, eliminating schools with missing data for any variable in the models.

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 1,189 schools. This group of schools became 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. This second sample contained mainly middle schools and accounted for 438 total schools in the Middle School sample.

Distributions of each grade-level sample. Initial analysis of the 2014-15 School Performance Grades showed very different distribution shapes when comparing the three groups of schools. The Elementary School sample demonstrated a slightly skewed left distribution with a higher proportion of schools receiving F grades than A grades. By contrast, the Middle School sample is skewed left with a higher proportion of schools receiving D and F grades than receiving A and B grades. Figure 3 summarizes the distributions by sample.

FIGURE 3 ABOUT HERE

After creating the samples based on the performance measures used for a school's grade, I removed any remaining charter schools from the sample because their characteristics differed from the general population of North Carolina public schools. Charter schools received performance grades and responded to the NC Teacher Working Conditions Survey, but these schools operate outside of Local Education Agencies and are not required to provide lunch and

transportation to students like traditional public schools. North Carolina also does not keep detailed teacher data on charter schools that will be necessary to assess actual teacher turnover.

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. Because a regression discontinuity looks at 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 aligned to one 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 two-level model guided the analysis for the first two research questions about teacher perceptions and immediate professional plans:

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

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01}(Treatment)_j + \gamma_{02}[\sum_{n=1}^{4}\left(SPG_{Score\,j} - SPG_{Cutoff}\right)^n] + \gamma_{03}W_j + \mu_{0j}$$

A two-level model accurately represented 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 β_{1j} coefficient in Level 1 and a β_{1j} equation in Level 2.

Y_{ij} represents the outcome of each scale mean for teacher *i* at school *j*. β_{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, γ_{00} represents the intercept and γ_{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 represented 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 γ_{02} coefficient controlled for 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 represented the distance of the school from the discontinuity. SPG_Score_j represented 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 includds linear, quadratic, cubic, and quartic terms of the distance of the running variable from the threshold. As shown in Table 2, for the "fail" vs. "pass" determination the value was 55, the minimum score to receive a C grade. Likewise, the value was 70 for the B/C cutoff and 40 for the D/F threshold. W_j and the corresponding γ_{03} coefficient represented school-level covariates. Finally, μ_{0j} signified the school-level mean for each outcome.

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 receive the same grade based on annual measures that can fluctuate from year to year. In theory, schools could manipulate 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.

Using the pretreatment outcome data, 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 elementary school analyses at the B/C, C/D, and D/F thresholds, I set the bandwidth to the 7, 10, and 15-point bandwidths.

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 presents 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 4 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.

TABLE 4 ABOUT HERE

Instead of specifying multiple potential treatments using combinations of one or both grades assigned, the main analysis relied on the most recent label. Thus, the treatment variable will correspond 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. I also ran a sensitivity analysis on the treatment by simply removing schools from the sample whose grades moved them from “fail” to “pass” or vice versa between the two school years. The results section provides more detail about this follow-up analysis.

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. These data come 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 and 38.8% of the middle school sample received School Performance Grades of D or F and had treatment value of 1. For the additional elementary school 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. Table 5 contains the sample sizes for the elementary analysis at various bandwidths around the grade cutoffs.

TABLE 5 ABOUT HERE

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. Specifying the running variable as a continuous score assigned unique values to each school, eliminating mass points occupied by multiple schools with 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 calculated the total number of students passing for each measure divided by the total number of students included for the measure. The 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.

Power analysis. I conducted a power analysis to determine whether a regression discontinuity was feasible for answering the research questions. I used assumptions of α = .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. I then wrote a function in R to calculate the full minimum detectable effect size (MDES) including the RD Design Effect multiplier. 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). The tab combined 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 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. Shown above, Table 5 lists the *n*-sizes associated with each cutoff for four proposed bandwidths for the elementary sample, the only sample with enough schools to conduct analyses at individual grade cutoffs. 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.

PowerUp does not calculate the value of ρ_{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 4, also neatly fit 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. Table 6 contains the calculated values for the RD Design Effect.

FIGURE 4 ABOUT HERE

TABLE 6 ABOUT HERE

Table 7 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. To obtain reliable estimates for the minimum detectable effect size (MDES), I used the 2016 data from the NC Teacher Working Conditions Survey to calculate the necessary parameters for each relevant item and scale for each sample.

All MDES values of .20 *SD* or lower are in bold. 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.

Table 7 contains 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. Thus, I assessed the impacts at different bandwidths for comparison.

TABLE 7 ABOUT HERE

Model covariates. The covariates in the model contributing most to increased power contain prior-year results for the outcome variables. 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.

I used publicly-available administrative data from the 2013-14 school year to model pretreatment student demographic characteristics. The model included the percentage of students classified as economically disadvantaged at the school level. NC DPI published the total number of students tested in each school who are economically disadvantaged and the total number of students tested. For these measures, I calculated a school percentage for the economically disadvantaged students in 2014 as a model covariate. I also added covariates for the percentage of students representing each racial group as coded by North Carolina. The model will represent the percentages of the seven race groups: American Indian, Asian, Hispanic, Black, White, Two or More Races, Pacific Islander, and White. The models contained the school-level percentage of

students from the first six race groups with the percentage of White students as the reference category.

The model also included demographic data about teachers. NC DPI published 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).

Principal turnover also has the potential to greatly impact teacher perceptions of their schools (Burkhauser, 2017). Thus, I included a dummy variable indicating if the school had a new principal since the previous observation. Data on principal turnover are not available in the public files on the DPI website. I created a variable for principal turnover using the teacher pay file from NCERDC to see whether the ID for the principal changed between 2014-15 and 2015-16, the year of the survey administration used for the outcome variables.

Constructing Outcome Variables

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 observations with one or more "don't know" responses to any item in the outcome variables from the data set.

Perceived support. I will combine three items from the North Carolina Teacher Working Conditions Survey that correspond to teachers' perceptions of support. Table 8 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 survey 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 followed the same coding scheme for all other outcomes in the first research question.

TABLE 8 ABOUT HERE

Perceived autonomy. The study will also measure 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 followed a four-point scale, and I calculated the mean for the three items for each teacher and the mean score for the three items for each teacher will serve 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. Unlike the previous two outcomes, this question was answered by teachers' responses to the single item, "State assessments accurately gauge students' understanding of standards." The survey item follows a four-point scale and the teacher-level value for this item served as 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 may also impact teachers' overall perceptions of their schools. I answered this question using teachers'

responses to the single item, "Overall, my school is a good place to work and learn." The teacher-level response to this item served as the dependent variable for the analysis of the *Good Place to Work/Learn* construct.

Intent to stay teaching at the same school. The final research question focused on the impact of a failing School Performance Grade on immediate professional plans. I measured this outcome using teachers' indicated intentions for their immediate professional plans on an item of the Teacher Working Conditions Survey. The item included six categorical responses, summarized with the state-level 2016 distribution of responses in Table 9. The second research question assessed planned retention based on this item, the percentage of teachers indicating they plan to continue teaching at their current school. Thus, I coded the teacher-level responses as 1 for indicating that teachers will "continue teaching at current school" and 0 for all other non-missing responses indicating a teacher intends to no longer teach in the school.

TABLE 9 ABOUT HERE

Results

Before running the HLM regression discontinuity models, I explored the trends of survey responses to the grade a school received. Table 10 contains a summary of the mean values for each item and each grade threshold. As hypothesized, the means had a positive correlation with four of five scales, meaning that a higher School Performance Grade was associated with a higher level of agreement for the *Support, Autonomy,* and *Good Place to Work/Learn* scales and items. A higher proportion of teachers in higher performing schools also indicated the intent to remain teaching in their school. The other scale, *Accuracy of State Assessments,* did not demonstrate a clear relationship between the grade a school received and the level of responses to the item on the Teacher Working Conditions survey.

TABLE 10 ABOUT HERE

Analytic Method

Determining impact. 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 as a linear, quadratic, cubic, and quartic term. 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.

To determine statistical significance with HLM, I built two models for each analysis: 1) a null model consisting of all covariates except the treatment variable, and 2) a full model that included the treatment indicator as the main effect of interest. Following the HLM estimation of the coefficients, I conducted an ANOVA test to determine whether the model with the treatment variable had a significantly different chi-square statistic from the null model.

Each model also includes weights that align to a triangular kernel. Thus, the model places more weight to observations close to the cutoff than those far away, considered a best practice in RD analysis (Lee & Lemieux, 2010). The *lmer* package did not include a built-in mechanism to apply a triangular kernel. Thus, I used the *kernelwts* function within the *rdd* R package 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 each HLM model by specifying the values under the *weights* parameter in each *lmer* model.

For the initial analysis of the impact of receiving a "pass" vs. "fail," I looked at all schools in the sample, setting the RD threshold to the C/D cutoff. All measured effect sizes were less than 0.1 SD at the discontinuity. As shown in Table 11, only one analysis showed statistical significance, with elementary teachers in failing schools indicating a 0.09 SD lower perception of support than teachers in schools receiving a passing grade, $X^2(1) = 4.09$, $p = .04$. However, this analysis undertook 10 comparisons, with two samples and five scales. When applying the Bonferroni adjustment for multiple comparisons this difference was no longer statistically significant. Thus, I could not reject the null hypothesis for any analysis; there existed no significant discontinuity at the pass/fail threshold for any of the research questions for elementary or middle schools.

TABLE 11 ABOUT HERE

The size of the elementary school sample allowed for a finer-grained analysis of individual grade cutoffs (i.e., B/C, C/D, and D/F). Table 2 contains the results for each scale and grade threshold for the elementary school sample. As shown in1 the table, there existed no statistically significant discontinuities at any of the three grade cutoffs for the elementary sample.

TABLE 12 ABOUT HERE

Sensitivity analysis. I conducted a follow-up sensitivity analysis for the pass vs. fail threshold for the elementary and middle school samples. The timing of the release of School Performance Grades occurred in February 2015 and September 2015, meaning that two different letter grades could potentially have impacted teacher perceptions. For this analysis, I removed all schools from the sample that received a different grade in 2013-14 from 2014-15. Although the removal of schools from the sample decreased statistical power, the analysis helped to rule out the possibility of two different grades contaminating the results. Table 13 summarizes the

findings from this updated analysis. As with the full 2014-15 sample, teachers at both elementary and middle schools that received the same grade for 2013-14 and 2014-15 showed no significant discontinuity at the pass/fail threshold.

TABLE 13 ABOUT HERE

Conclusions and Discussion

Although correlated with the School Performance Grade score, there was not significant evidence of a discontinuity at any of the cutpoints measured. That is for the pass/fail cutoff for elementary and middle schools and for the three individual grade cutoffs for elementary schools. This could be evidence that a) the grade label alone did not significantly impact teacher perceptions, b) the Teacher Working Conditions Survey is not designed to detect differences in teacher perceptions related to the letter grade assigned, or c) the impact of the grade assigned in September decayed in the six to seven months before teachers responded to the Working Conditions survey, d) the two grades assigned in 2013-14 and 2014-15 contaminated the reactions by misrepresenting a true cutoff. The following paragraphs offer some initial insight about what the results could mean and directions for future study to test each explanation presented in this paragraph.

For explanation (a), there are two potential directions for future investigation. One direction involves looking further into the past at teacher responses to the survey related to other accountability system changes. For example, the shift to Common Core standards and new tests in 2012-13 may have provided an initial exogenous shock that dampened the response to the letter grade in 2013-14 and 2014-15. Exploration of discontinuities in teacher responses to the 2014 NC Teacher Working Conditions survey would offer some insight if teacher perceptions and intent to stay in their schools decreased in response to this accountability shift. Given the correlation between teacher responses and raw student performance suggests that the change to

Common Core could represent an inflection point, rather than upon the application of the A-F labels. It is also possible that the two-decades-long presence of high-stakes accountability in the state already shifted teacher perceptions. The sequence of introduction of the labels after a long-established system, unlike the introduction of A-F grading in Florida in 1999, represents a different context in which the A-F label may not be as potent.

Explanation (b) also offers a plausible elucidation of the issue. Although I carefully chose items that aligned best to the changes associated with letter grades in theory, no item or scale asked directly about reactions to the A-F accountability system. This lack of measurement precision could explain the null impact findings. Collecting more targeted data related to the A-F label, as the path taken by the Smith and Imig (2017) survey of North Carolina principals, could improve measurement precision to see if the A-F may have an impact unobservable by the Working Conditions survey.

Explanation (c) explores the potential of decay in response to the label. Teachers may have had an initial response to the label when it was released in September 2015 that could have decayed by the time they took the survey in Spring 2016. The gap in timing between the release of School Performance Grades and administration of the Teacher Working Conditions survey every two years will always make this a limitation.

Finally, explanation (d) relates to the potential contamination of two different treatments. Given the data available, this alternative explanation was the only one I could test for this paper. As shown in the sensitivity results in Table 13, retaining only schools that received the same grade in each year prior to the 2016 survey also yielded no significant results in the pass vs. fail analysis. Given the samples and research questions I explored for this paper, I can safely rule out this final hypothesis as a potential explanation for results.

Completion of my dissertation research will continue to explore these possible explanations and tie their interpretation back to prior literature on school accountability and the A-F school grading policy. The final research will also explore the outcome of actual teacher retention as opposed to indicating an intent to stay on the survey. This research will also explore all outcomes for high schools.

In addition, it is possible that the A-F labels associated with school accountability may not affect teachers in their roles due to insulation from public press and criticism that school administrators must field. I also plan to conduct a follow-up analysis on survey responses from school administrators to see if the label potentially impacted their perceptions in their roles.

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

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.

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

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

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

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

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.

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.

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/

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

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

Mississippi Center for Public Policy. (2012, September 17). New A-F grades for 2011-2012. Retrieved from http://www.mspolicy.org/new-a-f-grades-for-2011-2012/

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

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

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 (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

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

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/

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.

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

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.

Tables and Figures

Table 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	2015-16	(Crain, 2016a; Crain, 2016b)
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.

Table 2
Conversion of School Performance Grade Score to Letter Grade

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

Table 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 Schools Only	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.

Table 4
School Performance Grade Transition Matrix from 2013-14 to 2014-15

		2014-15				
		A	B	C	D	F
2013-14	A	**2%**	1%	0%	0%	0%
	B	1%	**18%**	5%	0%	0%
	C	0%	5%	**34%**	5%	0%
	D	0%	0%	6%	**15%**	2%
	F	0%	0%	0%	2%	**3%**

Table 5
Sample Size of Schools at Each Cutoff for the Elementary Sample by Four Selected RD Bandwidths

	7 Point Bandwidth	10 Point Bandwidth	15 Point Bandwidth
A/B Cutoff	98	156	300
B/C Cutoff	407	566	811
C/D Cutoff	415	569	821
D/F Cutoff	167	227	361

Table 6
RD Design Effect Parameter Estimates at each Cutoff for Four Selected Bandwidths for the Elementary School Sample

	7 Point Bandwidth	10 Point Bandwidth	15 Point Bandwidth
A/B Cutoff	3.07	2.45	1.86
B/C Cutoff	3.80	3.67	3.33
C/D Cutoff	3.47	3.23	2.81
D/F Cutoff	2.41	1.79	1.42

Table 7
Minimum Detectable Effect Size (MDES) for the Pass/Fail Analysis by Teacher Perception Construct

	Elementary Schools	Middle Schools
Support	**0.09**	**0.13**
Autonomy	**0.10**	**0.15**
Accuracy of Accountability Measures	**0.08**	**0.12**
Overall Work Climate	**0.09**	**0.13**

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

Table 8
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 Accountability Measures	--	State assessments accurately gauge students' understanding of standards.
Overall Work Climate	--	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

Table 9
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

Table 10
Means by School Performance Grade and Sample

		2014-15 School Performance Grade				
Scale	Sample	A	B	C	D	F
Support	Elementary	3.43	3.28	3.03	2.80	2.58
	Middle	3.30	3.17	2.95	2.68	2.52
Autonomy	Elementary	3.15	3.19	3.08	2.99	2.82
	Middle	3.31	3.13	3.07	3.00	2.95
Accuracy of State Assessments	Elementary	2.32	2.35	2.31	2.35	2.38
	Middle	2.35	2.26	2.22	2.26	2.42
Good Place to Work/Learn	Elementary	3.37	3.38	3.25	3.06	2.80
	Middle	3.41	3.33	3.21	3.03	2.87
Intend to Stay Teaching at School	Elementary	0.88	0.87	0.83	0.75	0.65
	Middle	0.87	0.84	0.81	0.75	0.67

Table 11
Pass/Fail Cutoff Analysis

Scale	Sample	Effect Size (SD)
Support	Elementary	-0.09
	Middle	-0.05
Autonomy	Elementary	-0.02
	Middle	0.08
Accuracy of State Assessments	Elementary	-0.04
	Middle	-0.05
Good Place to Work/Learn	Elementary	-0.01
	Middle	-0.02
Intend to Stay Teaching at School	Elementary	-0.03
	Middle	0.02
Teacher Response Sample Size	Elementary	37,459
	Middle	16,387
School Sample Size	Elementary	1,189
	Middle	438

*Statistically significant at the $p < .05$ level.

Table 12
Elementary School Sample Effect Sizes at Various Cutpoints and Bandwidths

		Bandwidth		
Scale	Cutpoint	7 pt	10 pt	15 pt
Support	B/C	0.00	0.04	0.04
	C/D	0.08	0.03	-0.02
	D/F	0.07	0.04	-0.10
Autonomy	B/C	-0.03	-0.03	-0.02
	C/D	0.01	0.02	0.04
	D/F	0.07	-0.02	-0.13
Accuracy of State Assessments	B/C	0.03	-0.01	-0.02
	C/D	0.11	0.08	0.02
	D/F	0.02	0.04	0.00
Good Place to Work/Learn	B/C	0.01	-0.01	-0.04
	C/D	0.06	0.05	0.04
	D/F	0.08	0.01	-0.10
Intend to Stay Teaching at School	B/C	0.06	-0.02	-0.01
	C/D	0.02	0.02	0.02
	D/F	0.11	0.03	-0.03
Teacher Response Sample Size	B/C	12,899	17,724	26,005
	C/D	13,254	17,832	25,472
	D/F	4,924	6,720	10,905
School Sample Size	B/C	407	566	811
	C/D	415	569	821
	D/F	167	227	361

*Statistically significant at the $p < .05$ level.

Table 13
Pass/Fail Analysis with Schools Changing Grades Removed

Scale	Sample	Effect Size (SD)
Support	Elementary	-0.06
	Middle	-0.09
Autonomy	Elementary	0.02
	Middle	0.00
Accuracy of State Assessments	Elementary	0.03
	Middle	-0.02
Good Place to Work/Learn	Elementary	0.01
	Middle	-0.09
Intend to Stay Teaching at School	Elementary	-0.01
	Middle	0.00
Teacher Response Sample Size	Elementary	26,399
	Middle	12,694
School Sample Size	Elementary	825
	Middle	333

*Statistically significant at the $p < .05$ level.

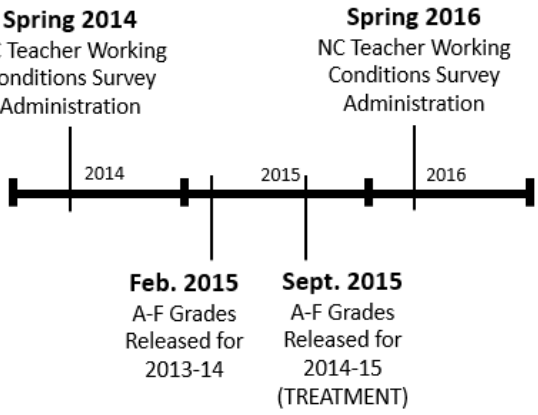


Figure 1. Timeline of events for study design.

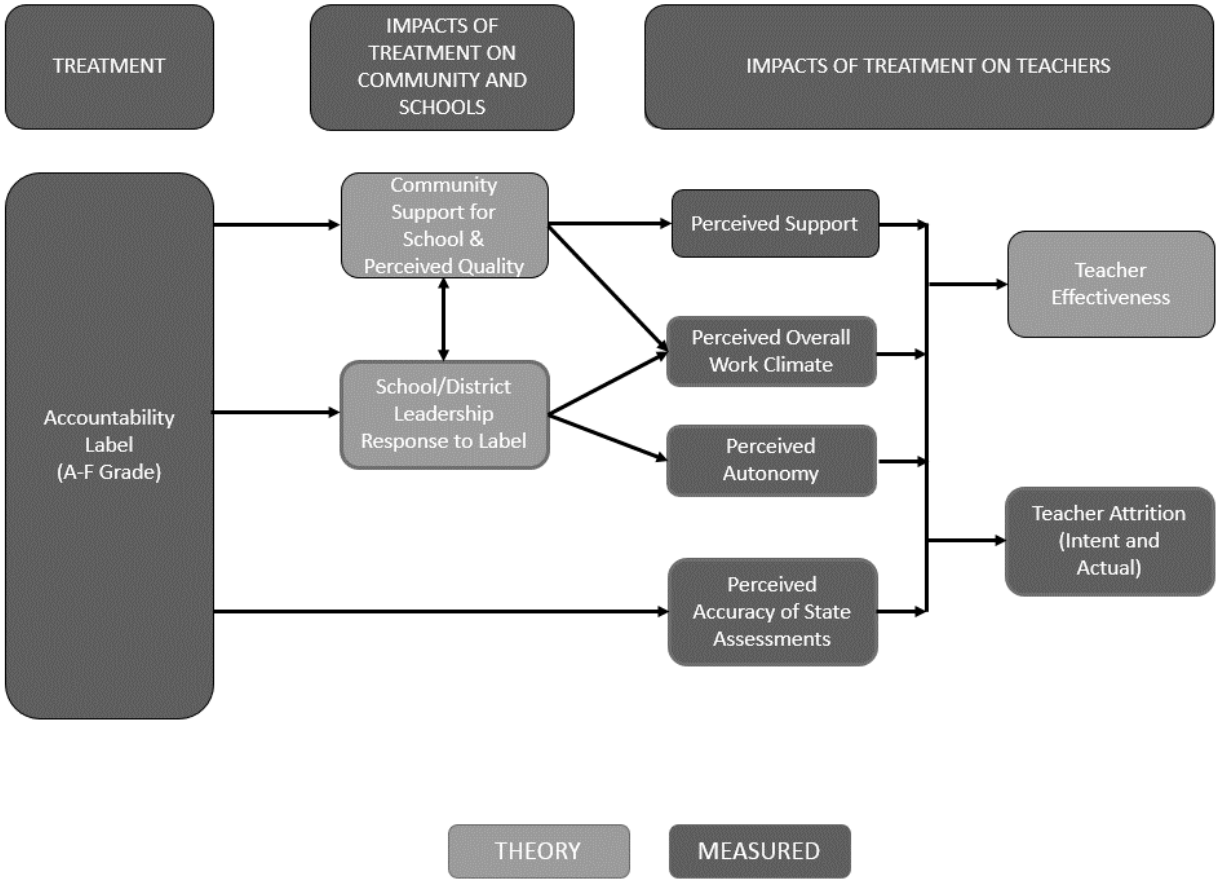
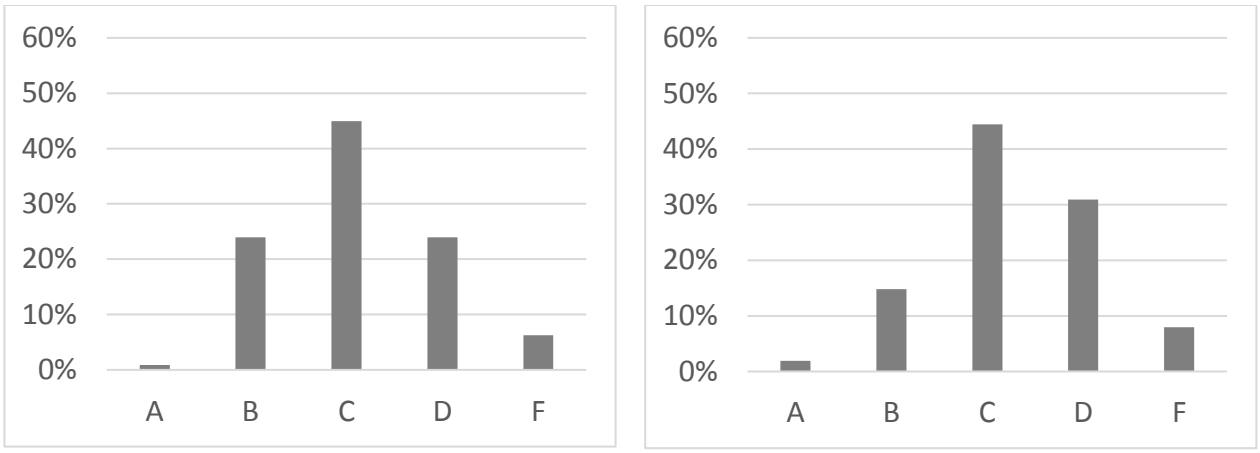


Figure 2. Conceptual framework.



Elementary Schools (*n* = 1,189) | Middle Schools (*n* = 438)

Figure 3. 2014-15 NC School Performance Grade distribution by sample conditions.

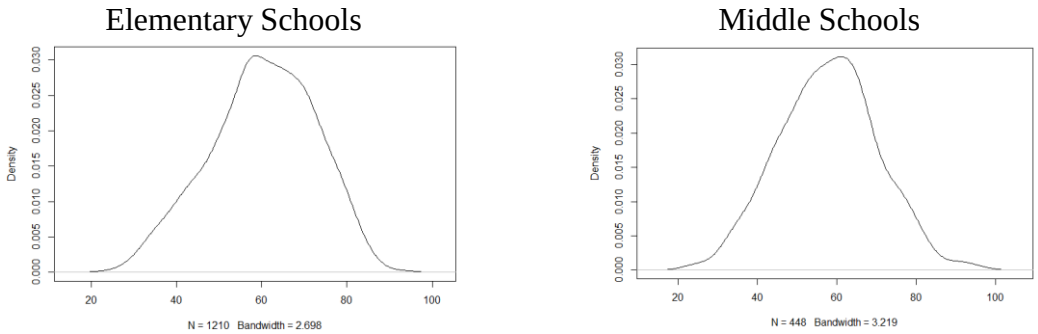


Figure 4. Distribution of the SPG score, the running variable for the RD.