Performance Standards and Employee Effort: Evidence from Teacher Absences

Seth Gershenson♦
School of Public Affairs
American University
4400 Massachusetts Avenue, NW.
Washington, DC 20016-8070.
Email: gershens@american.edu.
Phone: (202) 885-2687.
Fax: (202) 885-2347.

♦ The author is thankful for financial support from the W.E. Upjohn Institute's Early Career Research Grant program and an American University Faculty Research Support Grant. Opinions reflect those of the author and not necessarily those of the granting agencies. Conference participants at the 2014 Fall Meeting of the Association for Public Policy Analysis and Management provided helpful comments. Ashlyn Holeyfield, Stephen Holt, and Katie Vinopal provided excellent research assistance. Any remaining errors are my own.

Performance Standards and Employee Effort: Evidence from Teacher Absences

Abstract

The 2001 No Child Left Behind Act (NCLB) increased accountability pressure in U.S. public schools by threatening to impose sanctions on Title-1 schools that failed to make Adequate Yearly Progress (AYP) in consecutive years. Difference-in-difference estimates of the effect of failing AYP in the first year of NCLB on teacher effort in the subsequent year suggest that on average, teacher absences in North Carolina fell by about 10% and the probability of being absent 15 or more times fell by about 30%. Reductions in teacher absences were driven by within-teacher increases in effort and were larger among more effective teachers.

JEL Codes: J45, J48, J22, I2
Keywords: performance standards, employee effort, teacher absences, accountability, NCLB

1. Introduction

The public-sector accountability movement originated in the U.S. in the 1980s under the premise that performance can—and should—be measured in the public and non-profit sectors (Figlio & Kenny, 2009). Recent reforms in numerous areas of the public sector, including public education, have focused on using objective, observable measures of performance to reward efficiency and responsiveness and to hold units accountable (Heckman, Heinrich, & Smith, 2011). In the context of K-12 public education, high-stakes (i.e., evidence- or test-based) accountability policies aim to evaluate teachers, schools, or students on the basis of students' performance on standardized exams (Ladd, 1996). Such policies now play a prominent role in both state and federal U.S. education policy (Figlio & Loeb, 2011). The rationale for such policies is that attaching incentives to students' performance on standardized exams will alleviate the principal-agent problem inherent in the relationship between stakeholders and schools and improve student outcomes as a result (Figlio & Loeb, 2011). More generally, education economists argue that properly-aligned incentives can increase student achievement and decrease the costs of public education (e.g., Hanushek, 1994; Hoxby, 2007).

The proliferation of state-level accountability policies in the 1990s and the 2001 passage of the Federal No Child Left Behind Act (NCLB) reflect education policymakers' enthusiasm for such policies and have provided researchers with numerous contexts in which to investigate the impact of such policies on student achievement. Figlio and Loeb (2011), Figlio and Ladd (2008), and Hout and Elliot (2011) provide thorough reviews of this literature, which generally finds modest but statistically significant effects of about 0.1 to 0.3 test-score standard deviations (SD). For example, Carnoy and Loeb (2002) and Hanushek and Raymond (2005) exploit cross-state variation in the strength of high-stakes accountability policies and find positive effects on student

test scores. Similarly, Dee and Jacob (2011) compare the effect of NCLB on student achievement in states that had pre-existing high-stakes accountability systems similar to NCLB to NCLB's effect in states that did not, arguing that states without preexisting high-stakes accountability policies were "treated more intensely" by NCLB, and find that NCLB improved math test scores in "treated" states by about 0.2 SD.

However, critics of high-stakes accountability policies worry that these test-score gains are illusory and reflect strategic responses by schools rather than true learning gains. For example, high-stakes accountability policies' relatively narrow focus on standardized test scores, usually in math and reading, may cause teachers and schools to divert resources and instructional time away from non-tested topics and skills that are valued by stakeholders and important for students' long-run socioeconomic success (Baker et al., 2010). Indeed, some evidence suggests that teachers "teach to the test," as Jacob (2005) uses an interrupted time series research design to find an effect of a high-stakes accountability policy in Chicago Public Schools of about 0.3 SD on high-stakes tests but no effect on low-stakes National Assessment of Educational Progress (NAEP) test scores. Similarly, Reback, Rockoff, and Schwartz (2014) find evidence that NCLB accountability pressure caused schools to shift time away from non-tested subjects like science and social studies. There is also evidence of more nefarious unintended consequences of high-stakes accountability policies. For example, schools have prevented low-performing students from taking the standardized tests on which accountability policies are based either by reclassifying certain students as non-tested special-education students (Cullen & Reback, 2006) or suspending them on test days (Figlio, 2006). There is even evidence of outright teacher cheating (Jacob & Levitt, 2003).

As a result, the mechanisms through which high-stakes accountability policies affect students' academic achievement are not entirely understood, but have implications for the design of future education policies. Increased teacher effort is one potential mechanism through which high-stakes accountability policies might improve student achievement, as teachers are among the most important school-provided educational inputs (Ahn, 2013; Hanushek & Rivkin, 2010; Jacob, 2013). Teacher attendance is one measure of teacher effort, or teacher productivity, that affects student achievement (Ahn, 2013; Clotfelter et al., 2009; Das et al., 2007; Duflo, Hanna, & Ryan, 2012; Herrmann & Rockoff, 2012; Miller, Murnane, & Willet, 2008). Moreover, teacher attendance is positively correlated with both principals' ratings of teachers and teachers' value-added scores (Jacob & Walsh, 2011). Teacher absences are also costly in other ways: the substitute teachers necessitated by teacher absences are financially costly (Roza, 2007) and teacher absences create negative externalities, as Bradley, Green, and Leeves (2007) find that teachers' absences are affected by the attendance of their peers. The current study contributes to our understanding of the mechanisms through which consequential accountability policies affect student achievement and, more generally, how public-sector employees respond to the threat of sanctions associated with performance standards by examining how, if at all, failing to make Adequate Yearly Progress (AYP) in the first year of NCLB affected teacher absence rates.

The direction of the effect of high-stakes accountability policies on teacher absences is theoretically ambiguous, as such policies might affect teachers in two ways.[1] On the one hand, the incentives provided by such policies are intended to increase teacher effort (Ahn, 2013; Hansen, 2009). Viewing attendance as a type of employee effort, high-stakes accountability

[1] The direction of the effect of accountability policies on teacher turnover is similarly theoretically ambiguous (e.g., Clotfelter et al., 2004; Feng et al., 2010).

policies are hypothesized to decrease teacher absences.[2] Indeed, Jacob (2013) uses a difference-in-difference (DD) research design to show that a policy change in Chicago Public Schools that increased principals' ability to dismiss probationary teachers resulted in a significant decrease in probationary teachers' absences. On the other hand, the stress and pressure placed on teachers by high-stakes accountability policies might increase the psychic costs associated with teaching and lead to higher rates of teacher absences as a result (Clotfelter et al., 2009; Ose, 2005; Johansson & Palme, 1996). Which of these two hypothesized effects dominates is an empirical question that I address in the current study using statewide administrative longitudinal data on public primary school teachers in North Carolina. Specifically, I employ a DD-style identification strategy using data from the first two years of NCLB. The treatment is failing to make AYP in the first year of NCLB, which "turned on" the threat of sanctions for failing to make AYP in two consecutive years. The main results suggest that the threat of sanctions following failure to make AYP in the initial year of NCLB caused a robust, statistically significant decrease in teacher absences that cannot be explained by preexisting differential trends in treated schools nor by changes in the composition of treated schools' teaching staffs. These effects are larger among more effective teachers and are arguably practically significant, representing 10% declines in teacher absences.

The paper proceeds as follows. Section 2 describes the relevant institutional details regarding teacher absences and accountability policy in North Carolina. Section 3 introduces a theoretical model of teacher absences from which the econometric model is derived. Section 4 describes the data and section 5 presents the results. Section 6 concludes.

[2] See Ahn (2013), Hansen (2009), and Jacob (2013) for further justification of the use of teacher absences as a proxy for teacher effort and productivity.

2. Background

2.1 Teacher Absences in North Carolina

Teachers in North Carolina's public schools are permitted to take a limited number of absences per year, though not without incurring some personal costs. The current study focuses on two particular types of absences—sick and personal leave—as these are the most relevant sources of teacher absences during the school year (Ahn, 2013; Clotfelter et al., 2009). Teachers accrue sick leave at a rate of one day per month. For this reason, the cost of sick days is decreasing in teacher experience. Unused sick days can be redeemed at retirement for additional pension benefits, which makes using accumulated sick days costly to teachers. Teachers can also use more than their allotted number of sick days at a cost of $50 per absence and are similarly charged $50 per day of personal leave. Clotfelter et al. (2009) show that between 1995 and 2004 the average public-school classroom teacher in North Carolina took about seven sick days and one personal leave day per academic year. As expected, absences are slightly more common among more experienced teachers. The authors also examine the predictors of teacher absences, finding that absences are more common among teachers in elementary schools and schools serving low-income students. Teacher absences are less common among teachers who have a masters degree, graduated from a selective undergraduate institution, or are National Board Certified (Clotfelter et al., 2009).

2.2 Preexisting ABC Accountability Policy

North Carolina first implemented the ABC accountability policy at the start of the 1996-1997 academic year. Henceforth, academic years are referred to by the year of the spring semester. The ABC policy is notable in that it focuses on achievement growth in addition to

achievement levels, as schools are thought to have relatively more influence on the former. Achievement is measured by student performance on end-of-grade tests in math and reading administered each spring in grades 3 through 8. The tests are state mandated, aligned with state standards, criterion referenced, and vertically aligned. Schools that fail to make expected growth are labeled as either "no recognition" or "low performing," depending on whether or not at least 50% of students score at or above grade level. Schools that make expected growth are labeled as such, and schools that exceed expected growth by 10% or more are classified as "high growth." However, there are no direct sanctions for failing to meet growth requirements. Ahn (2013) carefully describes how these classifications are made.

Teachers in "high" and "expected" growth schools receive annual bonuses of $1,500 and $750, respectively. Ahn (2013) exploits the resulting discontinuity in award receipt to examine how teachers' effort responds to incentives. Intuitively, Ahn finds a U-shaped relationship between teacher effort and the probability of receiving a bonus, with highest effort occurring when bonus receipt is uncertain. Similarly, Clotfelter et al. (2004) point out that while ABC's focus on growth provides an arguably more valid measure of school performance, it is not perfect, as there may be numerous barriers to growth in low-performing schools. For this reason, teachers may attempt to transfer out of schools with low baseline achievement levels. The authors estimate the effect of ABC on teacher turnover using a DD strategy that compares turnover in low- and high-performing schools, before and after the implementation of ABC, and find that ABC increased turnover probabilities by about 25% in low-performing schools. However, given the current study's focus on the introduction of NCLB, it is worth stressing that the determinants of AYP and expected growth are mostly unrelated and nearly half of schools that make expected growth fail AYP, and vice versa (Ahn, 2013).

2.3 The 2001 No Child Left Behind Act (NCLB)

NCLB extended North Carolina's existing ABC accountability policy in three ways. First, it required that all schools make adequate yearly progress (AYP), which differed from existing ABC growth requirements by requiring that schools meet percent proficient, attendance, and test-participation thresholds both overall and for specific subgroups of the student population. Second, NCLB mandated that states publish "school report cards" containing information on schools' performance levels and AYP status. Third, NCLB mandated additional sanctions on Title-1 schools that failed to make AYP in consecutive years. The current study focuses on Title-1 schools, as such schools comprise the majority of North Carolina's public primary schools and the threat of sanctions is particularly salient. The identification strategy described below exploits the fact that Title-1 schools that failed to make AYP in 2003, the first year of NCLB, were under considerably more pressure in 2004 than Title-1 schools that made AYP in 2003. This idea is similar to that used by Chakrabarti (2014), who also considers 2003 a "pre-program" year, and Ahn and Vigdor (2014), who exploit variation in schools' AYP histories. The current study focuses on the first two years of NCLB for two reasons. First, data on teacher absences are unreliable after 2004 (Ahn, 2013). Second, the determination of AYP became more complex after 2004 due to NCLB waivers and "Safe Harbor" exemptions (e.g., Polikoff & Wrabel, 2013).

Title 1 was a component of the original ESEA that was reinstituted by NCLB, which provides federal funds to schools in proportion to the number of low-income students attending the school (Gordon, 2008). This money can be used to cover the cost of tutoring, after-school, and summer programs that reinforce the school's standard curriculum. When NCLB was first implemented, it mandated that Title-1 schools that failed to make AYP for two consecutive years

enter Program Improvement (PI). PI is a five-year process of steadily increasing consequences that culminates with the drastic restructuring of the school (e.g., the school is reinvented as a charter, taken over by the state, or replaces a majority of the staff). Therefore, the passage of NCLB placed pressure on all Title-1 schools and this pressure increased in severity in Title-1 schools that failed AYP in the first year of NCLB. Accordingly, DD estimates of the effect of failing AYP on teacher absences likely underestimate the total effect of NCLB on teacher absences, as the policy placed pressure on all schools in 2004, even those that did not fail AYP in 2003 (Fuller & Ladd, 2013).

3. Modelling Teacher Absences

3.1 Theoretical Model and Comparative Statics

Some teachers surely relish the opportunity to improve student performance and increase effort in response to accountability pressure, if for no other reason than to avoid the negative consequences of failing to make AYP (Figlio & Loeb, 2011). Not all teachers necessarily respond in this way, however, as high-stakes testing decreases teachers' classroom autonomy and sense of job security (Reback et al., 2014) and might increase teachers' stress levels (Barksdale-Ladd & Thomas, 2000; Daly & Chrispeels, 2005; Fuller & Ladd, 2013). Indeed, empirical evidence suggests that the exogenous increase in accountability pressure created by an unexpected policy change in Florida increased teacher turnover (Feng et al., 2010). On average, then, the net effect of high-stakes accountability policies on teacher effort, as measured by teacher absences, is theoretically ambiguous. Whether the increase in disutility associated with teaching, particularly among teachers in tested (high-stakes) grades and threatened (non-AYP) schools, outweighs the incentives to increase effort is an empirical question.

It is useful to formalize the intuitive argument made above by viewing employees' daily attendance (i.e., absence) decisions as daily labor-supply decisions. Assuming that teachers maximize expected utility when making such decisions, teachers' optimal behavior is governed by a simple "reservation wage" decision rule (Bradley et al., 2007; Gershenson, 2012). Formally, teacher *i* chooses to work on day *d* if the expected utility of working $\left(U_{id}^{W}\right)$ exceeds the expected utility of being absent $\left(U_{id}^{A}\right)$, and is absent otherwise. Assuming that these daily utilities contain additively-separable stochastic terms, the probability that teacher *i* is absent on day *d* is

$$\Pr\left(a_{id}=1\right)=\Pr\left(U_{id}^{A}\geq U_{id}^{W}\right)=G\left(U_{id}^{A}-U_{id}^{W}\right), \tag{1}$$

where *G* is the CDF of the distribution of the difference between the two stochastic components of daily utility, as in the standard random utility model (Cameron & Trivedi, 2005, p. 477).

In expectation, then, teacher *i*'s total annual absences (*A*) are simply

$$E\left(A_i\right)=\sum\nolimits_{d} G\left(U_{id}^{A}-U_{id}^{W}\right). \tag{2}$$

Assuming that U^W is decreasing in hours worked (*H*), the disutility associated with teaching (Ψ) is increasing in accountability pressure (*p*), and that U^A is decreasing in *p* due to the relationship between teacher absences and student achievement, straightforward application of the chain rule shows that the direction of the effect of *p* on teacher absences is ambiguous.[3] Formally,

$$\frac{\delta E\left(A\right)}{\delta p}=\sum\nolimits_{d} G'\left(\cdot\right)\left(\frac{\delta U^{A}}{\delta p}-\frac{\delta U^{W}}{\delta \psi}\cdot\frac{\delta \psi}{\delta p}\right), \tag{3}$$

where G' is strictly positive and the second term on the RHS of (3) is positive if and only if $\left|\frac{\delta U^{A}}{\delta p}\right|<\left|\frac{\delta U^{W}}{\delta \psi}\frac{\delta \psi}{\delta p}\right|$. Note that the sign of (3) is ambiguous even under the strong assumption that

[3] See Blundell & MaCurdy (1999) and Stern (1986) for descriptions of utility functions that incorporate the disutility of working and their corresponding labor supply functions. Two common examples are the Frisch and Stone-Geary functional forms.

teachers' preferences are constant across days; daily fluctuations in preferences provide an additional potential source of ambiguity. The empirical strategy for identifying the sign of (3) (i.e., how teacher absences respond to accountability pressure) is presented in section 3.2.

3.2 Econometric Model and Identification Strategy

The baseline analytic sample is restricted to the 2003 and 2004 academic years, as 2003 was the first year of NCLB and thus the first year that AYP was computed. Given the threatened sanctions associated with failing AYP in two consecutive years, 2004 was the first year that could possibly send a school to the PI program. Accordingly, the preferred model of teacher *i*'s absences (*A*, in levels), while in school *s* and year *t*, is a difference-in-difference (DD) specification similar to the baseline specification in Jacob (2013). Formally,

$$A_{ist} = \gamma d04_t + \tau d04_t \times Failed03_s + \beta X_{ist} + \theta_s + \varepsilon_{ist}, \quad (4)$$

where $d04$ is a binary indicator equal to one in 2004 and zero in 2003; $Failed03$ is a binary indicator equal to one if school *s* failed to make AYP in 2003 and zero otherwise; X is a vector of observed teacher and school characteristics including the teacher's National Board Certification, experience, educational attainment, selectivity of undergraduate institution, race, and gender and the school's total enrollment, fulltime equivalent teachers, student-teacher ratio, percent of enrollment eligible for free or reduced price lunch, and percent of enrollment that is non-white; θ is a school fixed effect (FE), and ε is an idiosyncratic error term. The elements of X and the school FE are consistent with the broad categories of theoretical determinants of employee absences identified by Ichino and Maggi (2000) and Bradley et al. (2007).[4] Conditioning on X

[4] Specifically, the categories are individual background, school background (i.e., locality or contextual effects), and group (social) interaction effects whereby group norms affect individual

and θ tend to improve the precision of estimates of τ, the effect of the increase in accountability pressure caused by the failure to make AYP in the inaugural year of NCLB, which is the primary parameter of interest. However, the main results are robust to omitting X and θ from (4).

Importantly, the school FE subsume the time-invariant *Failed*03 indicator and imply that comparisons are made within, as opposed to between, schools. Still, the possibility of pre-existing school-specific trends is a threat to identification, as schools that failed to make AYP in 2003 may have been on systematically different growth paths than those that did make AYP. I test the importance of this concern using an augmented event-study version of (4) using data dating back to 1997 that includes a full set of year×*Failed*03 interactions. Similarly, equation (4) can be augmented to include either teacher FE or teachers' lagged absences to control for the nonrandom sorting of teachers to schools. Finally, I test for heterogeneous effects by teacher characteristics by augmenting (4) to include a full set of interaction terms. The baseline models are estimated by OLS and standard errors are clustered at the school level.[5] A number of additional sensitivity analyses are conducted and discussed in section 5.

4. Data

The primary analyses are of primary (K-5) public school teachers who taught in Title-1 schools in North Carolina in 2003 and 2004, though the event-study analysis utilizes data going back to 1997. These longitudinal teacher-level administrative data are maintained and provided by the North Carolina Education Research Data Center (NCERDC).[6] The NCERDC data contain

---

absence rates. Ose (2005) also provides theoretical and empirical support for the hypothesis that workplace environments might affect employee absences.

[5] Teacher absences are not actually a count variable because the data contain fractions of absences, so negative binomial and poisson regressions are not appropriate.

[6] See http://www.childandfamilypolicy.duke.edu/project_detail.php?id=35.

administrative records on teachers' absences per pay period, race, gender, experience, National Board Certification, educational attainment, and undergraduate institution attended. The baseline analytic sample is restricted to teacher-years in which these variables, and school-level information on total enrollment, student demographics, and fulltime equivalent teachers are observed in the first two years of NCLB. There are 8,080 such teacher-years.

Following previous research on teacher absences in North Carolina (Ahn, 2013; Clotfelter et al., 2009), total annual teacher absences are calculated as the sum of absences coded as either sick or personal leave. The third broad class of absence, vacation leave, typically occurs during days that school is not in session and is thus excluded from the operationalized definition of teacher absences. Column (1) of table 1 reports summary statistics for the baseline analytic sample. The average teacher was absent almost nine times per year, which is consistent with previous research and largely driven by sick leave. The standard deviation of nearly ten indicates significant variation in teacher absences.

To better understand the variation in teachers' absences I decompose the total variation into within -school, -school year, and -teacher variation by estimating "within-unit" SD. I do so by computing the SD of the residuals from regressions of teacher absences on sets of school, school-by-year, and teacher FE, respectively. The within-school and within-school year SD reported in column 1 for the baseline analytic sample are quite close to the overall SD, indicating that nearly 97% of the total variation in teacher absences occurs within, as opposed to between, schools. Moreover, about 95% of the total variation occurs within school-years. Importantly, these results suggest that there is sufficient within-school and within-school year variation to identify the preferred school-FE specifications. The within-teacher SD is about half as large as the overall SD, indicating that about half of the total variation in teacher absences is approximately evenly

split between and within teachers. This suggests that teacher absences are somewhat "sticky," but again there is enough within-teacher variation to implement teacher FE estimators.

The analytic sample is approximately evenly split across the 2003 and 2004 school years, which were the first two years of NCLB, and across tested (grades 3 through 5) and non-tested (Kindergarten and grades 2 and 3) grades. About one quarter of teachers had earned a Masters degree and a trivial percentage had earned a doctorate. Teachers who reported less than a four-year degree are excluded from the sample. Five percent of teachers attended a selective undergraduate institution, where selectivity is coded as a dichotomous variable based on ratings from Barron's Profiles of American Colleges. Specifically, the selective indicator equals one if Barron's rated the institution as "most" or "highly" selective, and zero otherwise. Ten percent of teachers were National Board Certified and the average teacher had 14 years of experience teaching in North Carolina. The analytic sample is predominantly female and white.

I also test for heterogeneity by teacher effectiveness using value-added measures (VAMs). The VAMs are generated by using NCERDC student-teacher matched student-level administrative data on end-of grade math test scores to estimate standard lag-score value-added models. Accordingly, VAMs are available for teachers who taught at least two years in fourth or fifth grade.[7] Math scores are used because the literature on teacher effectiveness typically finds greater variation in teachers' effects on math than on reading (e.g., Gershenson, 2015; Hanushek & Rivkin, 2010). Importantly, I estimate VAMs for all teachers who taught in 2003 and 2004, not only those for whom absence data are available, and then record the quartile of the full distribution of teacher effectiveness in which teachers in the baseline analytic sample are located.

[7] The basic value-added model is similar to those estimated in Gershenson (2015) using similar NCERDC data. See Appendix A for more description of the value-added model and estimation.

Column (2) of table 1 reports summary statistics for the subsample of the baseline analytic sample for whom VAM scores are available. The VAM subsample summarized in column 2 closely resembles the baseline sample summarized in column 1, with two notable exceptions: the VAM subsample contains more tested-grade and male teachers. Intuitively, the former is due to VAM scores being unavailable for teachers in non-tested grades. Tested-grade teachers do not comprise the entire VAM subsample, however, because some teachers transitioned to a tested grade after being in a non-tested grade in 2004. Brummet et al. (2014) and Ost (2013) show that grade switching is fairly common among primary school teachers. The latter is due to men being more likely to teach in higher grades, which in the primary school context are also the tested grades.

5. Results

5.1 Main Results

Table 2 reports estimates of the baseline DD model shown in equation (4). Column 1 reports estimates of a simple DD specification that controls for neither teacher characteristics nor school FE. This coefficient estimate is equivalent to the double difference between average teacher absences in 2003 and 2004, and between schools that passed AYP in 2003 and schools that failed AYP in 2003. The resulting estimate of -1.25 is strongly statistically significant and suggests that on average, failing to make AYP in the initial year of NCLB decreased annual teacher absences by about 1.25 absences. Columns 2, 3, and 4 of table 2 add teacher covariates, school FE, and time-varying school covariates to the DD regression model, respectively, resulting in similarly sized treatment effect estimates. It is reassuring that the DD point estimate is robust to conditioning on teacher characteristics and time-varying school characteristics, as

this suggests that the estimates are not biased by underlying trends in the composition of schools' faculty or enrollments. The specification reported in column 4 of table 2 that conditions on school FE and teacher and school covariates is the preferred baseline specification.

The remainder of table 2 investigates the robustness of the baseline results to the functional form through which teacher absences enter the model. Because teacher absences are strictly nonnegative and zeros are exceedingly rare in the data, Column 5 of table 2 estimates the baseline specification using the natural log of total teacher absences as the dependent variable.[8] The DD estimate remains negative and statistically significant at the 5% confidence level. Specifically, the point estimate of -0.09 suggests that failing AYP is associated with approximately a 9% decrease in teacher absences.

Finally, following Jacob (2013), columns 6-9 of table 2 report estimates of binary outcome models in which the dependent variable equals one if the teacher was absent 15 or more times, and zero otherwise. Column 6 presents OLS estimates of a Linear Probability Model (LPM) that is otherwise equivalent to the preferred baseline specification. The LPM estimate suggests that, on average, failing AYP decreased the probability of a teacher being absent 15 or more times by three percentage points, and this effect is statistically significant at the 5% confidence level. From a base of 11%, this represents a 27% decrease in teachers who were absent 15 or more times per school year. Importantly, this result suggests that the pressure associated with failing AYP drastically changed some teachers' effort levels. Column 7 reports estimates of the same LPM for the restricted sample of schools that experienced variation in the "absent 15 or more times" indicator. It is reassuring that these estimates are nearly identical to

[8] Zeros are rare because many teachers have 0.5 annual absences, perhaps because of how an in-service professional development program was coded. In the log specification zeros were replaced with 0.5, the smallest non-zero value observed in the data.

those presented in column 6, as the FE logit (Chamberlain, 1980) estimates reported in column 8 of table 2 make the same sample restriction. The FE-logit coefficient reported in column 8 is negative and statistically significant at the 5% confidence level, but its magnitude cannot be directly compared to the LPM coefficients, nor can precise average partial effects (APE) be computed because values of the FE are unobserved. However, scaled coefficients that are rough approximations of APE, which can be compared to the LPM coefficients, can be computed using the product of the sample average of Pr(A > 15) (0.11) and one minus this probability (0.89) as an approximate scaling factor. The resulting scale factor of 0.098 implies an APE of about -0.034, which is in line with the LPM estimates reported above.

Overall, the results presented in table 2 provide consistent evidence that failing AYP in 2003 caused a significant decline in teacher absences (increase in effort) in 2004, regardless of how teacher absences are measured. Moreover, the DD estimates are robust to conditioning on observable teacher characteristics, school FE, and time-varying observable school characteristics, which suggests that the results are not driven by secular trends in the composition of schools' enrollments or teaching staffs. Nonetheless, the next section presents a series of sensitivity analyses and falsification tests that further probe the robustness of the results. To place the baseline estimates presented in table 2 in context, it is useful to compare them to the estimates of other interventions' effects on teacher absences. One relevant comparison is Jacob (2013), who estimates the effect of a change in policy in Chicago Public Schools during the early 2000s that decreased the job security of probationary teachers by increasing principals' ability to dismiss such teachers. Using a DD strategy similar to that employed in the current paper, the author finds that the policy change decreased teacher absences by about 1 absence per year, and decreased the probability that a teacher had 15 or more annual absences by about four percentage points. These

point estimates are remarkably similar to those reported in table 2, as are the estimated effects when converted to percentages. Similarly, Clotfelter et al. (2009) find that directly charging teachers $50 per absence would reduce average annual absence rates by about one full absence. Finally, reducing the number of teachers who are absent more than 15 times per school year has arguably practically significant effects on student achievement: 10 teacher absences reduce math achievement by about 0.02 math score SD, which is equivalent to replacing an average teacher with one from the bottom quintile of the effectiveness distribution (Clotfelter et al., 2009; Herrmann & Rockoff, 2013).

5.2 Sensitivity Analyses

Table 3 reports a variety of robustness checks and sensitivity analyses designed to test the baseline DD estimate's robustness to a variety of modeling choices, assumptions, and estimation strategies. Columns 1 and 2 estimate augmented versions of the preferred baseline specification that control for unobserved teacher heterogeneity by conditioning on lagged absences and teacher FE, respectively. The lagged-absences DD estimate in column 1 is slightly smaller than the preferred baseline estimate, though it remains larger than one in magnitude and is significant at the 5% confidence level.[9] The two-way school-teacher FE DD point estimate reported in column 2 is nearly identical to the lagged-absences estimate shown in column 1, though imprecisely estimated.[10] The imprecision of the two-way FE estimate is unsurprising, as the identifying variation in this model comes from the small number of teachers who changed schools between the 2003 and 2004 school years. Nonetheless, the similarity between the lagged-

[9] The sample size in the lagged specification is smaller than in the baseline because data on 2002 absences are missing for many teachers.

[10] The two-way FE specification is estimated using the estimator proposed by Mittag (2012).

absences and two-way FE estimates is reassuring given the bracketing property of these two estimators (Angrist & Pischke, 2009, p. 245). Together, the similarities between the baseline estimate of -1.37 and the estimated effects in columns 1 and 2 of table 3 suggest that the baseline DD estimate is not biased by endogenous sorting of teachers into schools. Moreover, the teacher-FE estimate in column 3 suggests that the results are driven by changes in teacher behavior, as opposed to changes in the composition of "treated" schools' teaching staffs. The estimates in columns 3 and 4 of table 3 further investigate this issue.

To provide further evidence that the accountability pressure associated with failing AYP caused a change in teacher behavior, as opposed to merely changing the composition of the teaching force in such schools, column 3 of table 3 estimates the baseline model on a restricted sample that excludes teachers who changed schools between 2003 and 2004. The resulting estimate is nearly identical to the preferred baseline estimate, again indicating that the threat of sanctions caused individual teachers to increase effort. Column 4 estimates a version of the preferred baseline model that replaces the school FE with teacher FE, using the same restricted sample. School FE are redundant in this specification, as the sample is restricted to teachers who did not change schools. Once again, the point estimate decreases slightly but remains larger than one in magnitude and is significant at the 5% confidence level. The estimates reported in columns 3 and 4 of table 3 are consistent with the lagged-absence and two-way FE estimates reported in columns 1 and 2 and provide additional evidence that the baseline DD estimate is capturing the behavioral response of individual teachers and not changes in schools' faculties.

Another potential concern is that the baseline estimates are influenced by the behavior of outliers who are absent at extremely high rates, as high rates of absence are more likely to be associated with health issues than shirking (Jacob, 2013). Accordingly, column 5 of table 3

reports estimates of the preferred baseline model on a restricted sample that excludes teachers who were absent 50 or more times in a given year.[11] The resulting point estimate falls slightly, to about negative one, but remains strongly statistically significant and is actually more precisely estimated than the preferred baseline estimate. Again, this result is consistent with the notion that the threat of NCLB sanctions caused a behavioral response in teachers' effort provisions.

Finally, the remaining columns of table 3 test the parallel slopes assumption, which is the crucial identifying assumption required for consistency of the DD estimator. Specifically, the DD estimates require that schools that failed AYP in 2003 were not already experiencing different trends in teacher absences than were schools that made AYP in 2003. Columns 6 and 7 do so indirectly in the spirit of regression discontinuity (RD) designs that exploit schools' "closeness" to the 2003 AYP margin (e.g., Ahn & Vigdor, 2014; Chakrabarti, 2014). Following Ahn and Vigdor (2014), I create a school-specific running variable (c) that equals the minimum of the differences between the actual 2003 proficiency rates and the 2003 thresholds for each relevant subgroup and the school as a whole.[12] According to the rule, when c is ≥ 0, the school should make AYP, and fail to make AYP otherwise. I use the variable c in two ways. First, in the spirit of the "discontinuity sample" sensitivity analysis advocated by Angrist & Pischke (2009, p. 257), I use c to restrict the baseline sample to schools that were "close" to the AYP margin. Column 6 of table 3 reports estimates of the baseline DD specification using a sample restricted to schools in the 10^{th}-90^{th} interpercentile range of c.[13] The resulting DD estimate is slightly smaller than the preferred baseline estimate but remains larger than one in magnitude and statistically significant at the 5% confidence level, suggesting that the main results are not driven by extremely high- or

[11] Qualitatively similar results are obtained using lower cutoffs (e.g., < 40 absences).
[12] See Appendix B for a precise description of how c is computed and RD estimation.
[13] Qualitatively similar results are obtained using smaller inter-percentile ranges (e.g., 25^{th}-75^{th}).

low-performing schools. Second, I use *c* to directly estimate a "simple" fuzzy-RD by 2SLS (Angrist & Pischke, 2009, p. 261) using 2004 data. The RD estimate is reported in column 7 and is similar in magnitude to both the baseline and "discontinuity sample" DD estimates, though it is imprecisely estimated.[14] Together, these RD-flavored sensitivity analyses suggest that the main results are not driven by schools in the tails of the "AYP likelihood" distribution, and highlight the importance of conditioning on school FE in the baseline DD specification.

Columns 8 and 9 conclude the series of robustness checks by directly addressing the parallel slopes assumption. Both analyses utilize data dating back to 1997, which is the omitted reference year. Column 8 estimates an augmented version of the preferred baseline model that conditions on school-specific quadratic time trends. The point estimate remains larger than one in magnitude and statistically significant at 5% confidence, which suggests that the baseline estimates are not the result of pre-existing differential trends in treated schools. Rather, teachers in schools that failed AYP in 2003 experienced, on average, a negative 1.65 absence deviation from trend in the second year of NCLB.[15] Similarly, column 9 reports estimates of an event-study specification that interacts a series of year indicators with the "failed AYP in 2003" indicator, which allows for a direct test for the presence of preexisting differential trends in teacher absences in the schools that ultimately failed AYP in 2003. It is extremely reassuring, particularly given the imprecision of the RD estimate, that the pre-2004 placebo effects (interaction terms) are neither individually nor jointly statistically significant. Moreover, the placebo effects are uniformly small in magnitude (< 0.30) and display no clear patterns or trends,

[14] Alternative RD specifications and bandwidths yield similarly imprecise, negative point estimates of similar magnitude. The simple RD estimate reported here is preferred because of its relative transparency and good finite sample properties (Angrist & Psichke, 2009, p. 261).
[15] A linear school trend specification yields a point estimate of the treatment effect that is nearly identical to the baseline DD estimate (1.37) and strongly statistically significant.

as half are positive, half are negative, and the magnitudes oscillate from year to year. In sum, the robustness checks and sensitivity analyses reported in table 3 provide consistent evidence of a statistically significant, negative, arguably causal effect of the threat of sanctions associated with failing AYP in the first year of NCLB on teacher effort in the subsequent year, of slightly more than one absence per year. In particular, the last two columns of table 3 provide strong evidence that the "parallel slopes" identifying assumption of the DD estimator is not violated, and hence that the DD estimates can be given a causal interpretation. The following section investigates whether these effects vary by observed teacher qualifications, teaching in a high-stakes (tested) grade, or teaching effectiveness.

5.3 Heterogeneous Treatment Effects

There are a number of reasons why the increased accountability pressure associated with failing AYP in the first year of NCLB might elicit varied responses from different types of teachers. For example, teachers in high-stakes (i.e., tested) grades may feel greater pressure to increase student achievement on the end-of-grade tests used to compute proficiency rates (Fuller & Ladd, 2013). In the analytic sample, grades 3 through 5 are tested and kindergarten, first grade, and second grade are not. The effect might also vary by observable teacher characteristics such as educational attainment, National Board Certification (NBC), experience, selectivity of undergraduate institution, and gender. Rationale for why the effect might vary by these characteristics is as follows. The first three characteristics all affect teacher pay scales in North Carolina. Because the cost of using sick days is loosely tied to earnings, and more strongly to experience, it is plausible that responses to increased accountability pressure vary by these characteristics. More generally, the training and experiences associated with these

characteristics, and with attending a selective undergraduate institution, might make some teachers more resilient to changes in the work environment. Lastly, regarding gender, it is hypothesized that women are more sensitive to changes in accountability pressure. For example, previous research shows that changes in public sector sick leave and monitoring policies have larger effects on women than on men (De Paola et al., 2014), which could be attributable to either differences in home responsibilities or psychological differences (Bertrand, 2011).

Column 1 of table 4 reports estimates of an augmented version of the baseline model that fully interacts the year, failed AYP, and year×failed AYP terms with these observable teacher characteristics.[16] Because teacher characteristics vary within school-years, it is now possible to replace the school FE with school-by-year FE. These estimates are reported in column 2 of table 4. The school FE and school-by-year FE estimates are similar to one another, which is reassuring and again suggests that the main results are not driven by school-specific trends. Only the selective undergraduate institution interaction terms are even marginally statistically significant, and as a whole the interaction terms are jointly insignificant. Still, while insignificant at traditional confidence levels, the tested-grade interaction effect of about one is interesting, as it suggests that the decrease in teacher absences documented to this point is stronger among teachers in low-stakes, non-tested grades. One possible explanation of this suggestive result is that the increased stress associated with teaching in a high-stress environment (e.g., Barksdale-

[16] Models that include one source of heterogeneity at a time yield qualitatively similar results. Experience and the corresponding interaction terms are modeled linearly, as the non-parametric estimates reported in Wiswall (2013) suggest that the returns to experience are approximately linear. However, the finding of no differential effects by experience is robust to modeling the interaction effect as a quadratic or cubic function in experience, and to instead using a "new teacher" binary indicator.

Ladd & Thomas, 2000; Daly & Chrispeels, 2005; Fuller & Ladd, 2013), which is associated with employee absences in other settings (Ose, 2005), is particularly acute in tested grades.

The selective undergraduate institution interaction effect is relatively large in magnitude and significant at the 10% confidence level, suggesting that teachers who attended selective undergraduate institutions experienced even larger decreases in absences (increases in effort) in response to failing AYP in 2003. Taken at face value, the average decrease of more than five annual absences represents a more than 50% decline in absences among such teachers. We can only speculate as to the reason for this apparent difference, which could be attributable to the higher levels of cognitive and non-cognitive skills associated with selection into, and accumulated at, selective undergraduate institutions. Regardless of the underlying cause, this result had potential implications for education policy and student achievement, as the selectivity of teachers' undergraduate institutions is weakly positively associated with student achievement gains (Boyd et al., 2008; Clotfelter et al., 2007; Rockoff et al., 2011). Finally, the statistically insignificant but positive male interaction effect suggests that women were marginally more affected by the threat of sanctions than men, which is consistent with De Paola et al. (2014).

Still, what might be of most interest to education policy makers is how teachers' effort levels respond to the threat of sanctions across the distribution of teacher effectiveness. It is well documented that the observable teacher qualifications discussed above explain, at best, only a small percentage of the total variation across classrooms in student achievement gains (e.g., Rockoff et al., 2011). Accordingly, as described in section 4 and Appendix A, I directly estimate teachers' math value-added measures (VAMs) using matched student-teacher data. I then create a set of categorical indicators that identify which quartile of the VAM distribution each teacher falls in and once again estimate augmented versions of the baseline model that fully interact the

year, failed AYP, and year×failed AYP terms with these categorical indicators of teacher effectiveness. Of course, VAMs are only available for teachers who taught in a tested grade in at least two years, so the VAM interaction specification can only be estimated on a subset of the baseline analytic sample.

Columns 3-5 of table 4 report three sets of estimates that rely on the subsample of teachers for whom VAM scores could be computed. Column 3 reports estimates of the preferred baseline model, which contains no interaction terms, to provide context. The DD point estimate is negative, slightly smaller than the preferred baseline estimate, and imprecisely estimated. It is unsurprising that the point estimate is smaller, as tested-grade teachers comprise the majority of the subsample and the interaction specifications reported in columns 1 and 2 of table 4 suggest that the effect was smaller among tested-grade teachers. Similarly, the estimate's imprecision may owe to the substantially reduced sample size. Columns 4 and 5 of table 4 report estimates of the VAM-interaction models that condition on school and school-by-year FE, respectively. Like in the teacher-qualification interaction models, the two VAM interaction models yield qualitatively similar results, though the interaction effects in the school-by-year FE specification are almost twice as large. Interestingly, the VAM interactions suggest that more effective teachers responded more strongly to the threat of sanctions than teachers in the bottom quartile of the effectiveness distribution. Specifically, these results suggest that more effective teachers decreased their absences by about 3 to 6 absences more than their less effective counterparts, which represents an arguably practically significant increase in effort. This is consistent with the finding of larger decreases in absences among teachers who attended selective undergraduate institutions and suggests that it is more effective teachers who are either able or willing to increase effort in response to the pressures associated with consequential accountability policies.

6. Conclusions

This paper estimates the effect of the threat of sanctions tied to failing to meet performance standards on employee effort in the public sector. The analysis exploits the fact that in the second year of NCLB, schools that failed to make AYP in the first year were at risk of facing the sanctions associated with failing to make AYP in two consecutive years. The results suggest that in year two of NCLB, teachers in schools that failed to make AYP in year one significantly reduced their annual absences by about 10%, or a little more than one absence per year. Similarly, the probability that a teacher was absent 15 or more times fell by 27%, or three percentage points, in such schools. The effect of NCLB's performance standards on teacher effort was concentrated among more effective teachers and the effect was mostly due to within-teacher changes in effort, as opposed to compositional changes in schools' teaching staffs.

These effects are arguably practically significant, as research on the harm associated with teacher absences in North Carolina finds that a one SD increase in teacher absences is associated with a decrease in math achievement of about 0.02 test-score SD, or 20% of the effect of a one SD increase in teacher effectiveness (Clotfelter et al., 2009; Herrmann & Rockoff, 2012). The estimated effects of performance standards on teacher effort are consistent with previous research on the malleability of teacher effort, as Ahn (2013) and Jacob (2013) find evidence that teacher effort, as measured by teacher absences, responds to incentives. Indeed, the magnitudes of the estimated effects in the current paper are similar to those of the estimated effects of a policy change in Chicago that granted principals the discretion to dismiss probationary teachers (Jacob, 2013). Finally, it is worth noting that the estimates reported here likely underestimate the total effect of NCLB on teacher absences and teacher effort, as the policy placed pressure on all schools, including those that made AYP in the first year (Fuller & Ladd, 2013).

The results of the current study have at least three implications for education policy. First, that teacher absences fell in response to increased accountability pressure suggests that one mechanism through which consequential accountability policies affect student achievement is through increased teacher effort. Second, these results contribute to the growing body of evidence that teacher effort, as measured by absences, responds to both school- and individual-level incentives. Finally, the heterogeneity in teachers' responses to the threatened sanctions associated with failing AYP in the first year of NCLB suggest potential benefits to policy designs and teacher training programs that account for such differences. For example, to the extent that teachers in tested and non-tested grades responded differently to the threat of sanctions, standard labor-economic theory suggests that if jobs in tested grades are more stressful, such jobs can pay compensating differentials. The wage differentials need not be monetary and could instead be provided in the form of additional planning periods, teaching aids, or professional development. Similarly, that the increase in effort was particularly strong among more effective teachers suggests that providing additional support to less effective teachers may be helpful, particularly for teachers and schools subject to increased accountability pressure.

Appendix A: Value Added Measures (VAMs) of Teacher Effectiveness

Following Gershenson (2015) and Jackson (2013), value-added measures of teacher effectiveness are generated by value-added models of the form

$$y_{ijgst} = \alpha y_{i,t-1} + \boldsymbol{\beta}\mathbf{x}_{it} + \boldsymbol{\gamma}\mathbf{c}_{-i,\,jgst} + \theta_j + \pi_g + \omega_{st} + u_{ijgst}, \qquad \text{(A.1)}$$

where *i*, *j*, *g*, *s*, and *t* index students, teachers, grades, schools, and years, respectively; *y* is performance on the end-of-grade math test, which is standardized by grade and year to have mean zero and standard deviation one (Ballou, 2009); **x** is a vector of observed student

characteristics including race, gender, poverty status, special education, and English language proficiency; **c** is a vector of classroom characteristics including class size, class composition, and the average of student *i*'s classmates' lagged achievement (peer effects); θ, π, and ω are teacher, grade, and school-by-year fixed effects (FE), respectively; and *u* is an idiosyncratic error term.

Equation (B.1) is estimated by Ordinary Least Squares for two reasons. First, Guarino, Reckase, and Wooldridge (2015) find OLS to be the most robust estimator to a variety of potential student-teacher assignment scenarios. This is potentially important, as Rothstein (2010) finds evidence of non-random sorting in North Carolina. Second, Chetty et al. (2014) find that most sorting of students to teachers is based on lagged test scores and that conditioning on lagged test scores alone yields estimated teacher effects with near-zero bias. Similarly, Kane and Staiger (2008) find that controlling for lagged test scores yields unbiased estimates of teacher effects and that controlling for classroom characteristics (i.e., the vector **c**) improves the precision of estimated teacher effects.

Equation (B.1) is estimated using student-level data on more than 1,112,000 fourth and fifth grade students and more than 32,000 fourth and fifth grade teachers between the years 2004 and 2010, as third-grade and 2003 data are lost in the creation of lag scores. Note that a VAM can be estimated for a teacher who did not teach in a tested grade in 2004, so long as the teacher did teach in a tested grade in at least two years during this time span. The estimated teacher effects are strongly jointly significant (F = 4.5) and the standard deviation of estimated teacher effects is 0.48. Estimated teacher effects range from -3.38 to 2.13. Quartiles of the full distribution of estimated teacher effects are -0.28, 0.06, and 0.29.

Appendix B: Regression Discontinuity (RD) Sensitivity Analysis

The RD sensitivity analyses reported in columns 7 and 8 of table 3 rely on a school-specific running variable (c) that measures the school's closeness to making AYP in 2003. In 2003, there were no waivers or "safe harbor" exemptions. In North Carolina, schools must make AYP both overall and in student subgroups that contain 40 or more students. Subgroups are Asian/Pacific Islander, Black, Hispanic, Multi-racial, Native American, White, Economically Disadvantaged, Limited English Proficient, and Students with Disability. If any subgroup's proficiency rate does not meet the threshold defined by the state, the school fails to make AYP. North Carolina's primary school math and reading proficiency thresholds in 2003 were 74.6% and 68.9%, respectively. Thus, following Ahn and Vigdor (2014), I define c as the minimum within-school difference between the actual proficiency rate of each subgroup and the state's AYP threshold. Proficiency rates are taken from data compiled by Reback et al. (2014), which the authors graciously make publicly available here: http://www8.gsb.columbia.edu/nclb/.

After creating c, I estimate a "simple" fuzzy-RD design by 2SLS (Angrist & Pischke, 2009, p. 261) using all 2004 data on teacher absences. Specifically, the first stage is

$$Failed_{is} = \pi_0 + \pi_1 c_s + \pi_2 X_{is} + \pi_3 T_s + \xi_{is}, \tag{B.1}$$

where T is a binary indicator equal to one if $c < 0$, and zero otherwise. The first stage results provide no evidence of a weak instruments problem. The second stage is then

$$A_{is} = \beta_0 + \beta_1 c_s + \beta_2 X_{is} + \tau Failed03_s + \varepsilon_{is}. \tag{B.2}$$

I report this "simple" fuzzy-RD estimate in table 3 because it is relatively transparent and has good finite sample properties (Angrist & Pischke, 2009, p. 261). However, a variety of alternative specifications were considered: including polynomials in c, allowing the slope to vary for positive and negative values of c, excluding X from the model, conditioning on school-district

FE, and trimming schools with high and low values of c from the analytic sample. These alternative specifications generally yield estimates of τ in range of -0.5 to -2.5, none of which are statistically significant at traditional confidence levels, regardless of whether the standard errors are clustered at the school level or not clustered at all.

References

Ahn, T. (2013). The missing link: Estimating the impact of incentives on teacher effort and instructional effectiveness using teacher accountability legislation data. *Journal of Human Capital*, 7(3), 230-273.

Ahn, T., & Vigdor, J. (2014). The impact of No Child Left Behind's accountability sanctions on school performance: regression discontinuity evidence from North Carolina. NBER Working Paper No. w20511

Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Ballou, D. (2009). Test scaling and value-added measurement. *Education Finance and Policy*, 4(4), 351-383.

Barksdale-Ladd, M. A., & Thomas, K. F. (2000). What's at Stake in High-Stakes Testing Teachers and Parents Speak Out. *Journal of Teacher Education*, 51(5), 384-397.

Bertrand, M. (2011). New perspectives on gender. *Handbook of Labor Economics*, 4, 1543-1590.

Blundell, R., & MaCurdy, T. (1999). Labor supply: A review of alternative approaches. In (eds.), *Handbook of Labor Economics*, vol. 3A, (pp. 1559-1695) Amsterdam: North Holland.

Boyd, D., Lankford, H., Loeb, S., Rockoff, J., & Wyckoff, J. (2008). The narrowing gap in New York City teacher qualifications and its implications for student achievement in high-poverty schools. *Journal of Policy Analysis and Management*, *27*(4), 793-818.

Bradley, S., Green, C., & Leeves, G. (2007). Worker absence and shirking: Evidence from matched teacher-school data. *Labour Economics*, *14*(3), 319-334.

Brummet, Q., Gershenson, S., & Hayes, M. (2013). The distribution of teachers' grade-level reassignments: Evidence from Michigan. Mimeo, American University.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: methods and applications*. Cambridge university press.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational evaluation and policy analysis*, *24*(4), 305-331.

Chakrabarti, R. (2014). Incentives and responses under *No Child Left Behind*: Credible threats and the role of competition. *Journal of Public Economics*, 110, 124-146.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *The Review of Economic Studies*, 47(1), 225-238.

Chetty, R., J.N. Friedman, and J.E. Rockoff. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, 104(9), 2593-2632.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, *26*(6), 673-682.

Clotfelter, C. F., Ladd, H., & Vigdor, J. (2009). Are teacher absences worth worrying about in the U.S.? *Education Finance and Policy*, 4(2), 115–149.

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.

Cullen, J. B., & Reback, R. (2006). Tinkering toward accolades: School gaming under a performance accountability system. In T. Gronberg & D. Jansen (Eds.), *Improving school accountability: Check-ups or choice?* (Advances in Applied Microeconomics, Vol. 14, pp. 1-34). Amsterdam: JAI Press.

Daly, A. J., & Chrispeels, J. (2005). From problem to possibility: Leadership for implementing and deepening the process of effective schools. *Journal for Effective Schools*, 4, 7-25.

Das, J., Dercon, S., Habyarimana, J., & Krishnan, P. (2007). Teacher shocks and student learning: Evidence from Zambia. *Journal of Human Resources*, 42(4), 820–862.

Dee, T. S., & Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management*, 30(3), 418-446.

De Paola, M., Scoppa, V., & Pupo, V. (2014). Absenteeism in the Italian Public Sector: The effects of changes in sick leave policy. *Journal of Labor Economics*, *32*(2), 337-360.

Duflo, E., Hanna, R., & S. P. Ryan. (2012). Incentives Work: Getting Teachers to Come to School. *American Economic Review*, 102(4), 1241-78.

Feng, L., D. N. Figlio, & Sass, T. (2010). School accountability and teacher mobility. NBER Working Paper No. 16070.

Figlio, D. N. (2006). Testing, crime and punishment. *Journal of Public Economics*, *90*(4), 837-851.

Figlio, D. N., & Kenny, L. W. (2009). Public sector performance measurement and stakeholder support. *Journal of Public Economics*, *93*(9), 1069-1077.

Figlio, D. N., & Ladd, H. F. (2008). School accountability and student achievement. *Handbook of Research in Education Finance and Policy*, pp. 166-182.

Figlio, D., & Loeb, S. (2011). School accountability. In E. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the Economics of Education*, vol. 3, (pp. 383-421). Amsterdam: North Holland.

Fuller, S. C., & Ladd, H. F. (2013). School-based accountability and the distribution of teacher quality across grades in elementary school. *Education Finance and Policy*, 8(4), 528-559.

Gershenson, S. (2012). How do substitute teachers substitute? An empirical study of substitute-teacher labor supply. *Economics of Education Review*, 31(4), 410-430.

Gershenson, S. (2015). Linking teacher quality, student attendance, and student achievement. *Education Finance and Policy*. In Press.

Gordon, N. (2008). The changing federal role in education finance and governance. *Handbook of Research in Education Finance and Policy*, 295-313.

Guarino, C. M., M. D. Reckase, and J. M. Wooldridge. (2015). Can value-added measures of teacher performance be trusted? *Education Finance and Policy.* In Press.

Hansen, M. (2009). How career concerns influence public workers' effort: Evidence from the teacher labor market. Calder Working Paper No. 40.

Hanushek, E. A. (1994). *Making schools work: Improving performance and controlling costs*. Brookings Institution Press.

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, *24*(2), 297-327.

Hanushek, E. A., and S.G. Rivkin. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review* 100(2), 267-271.

Heckman, J.J., C. J. Heinrich, & J. Smith. (2011). Performance standards and the potential to improve government performance. In Heckman, J. J. (Ed.) *The performance of performance standards*, pp. 1-14. WE Upjohn Institute: Kalamazoo, MI.

Herrmann, M. A., & Rockoff, J. E. (2012). Worker Absence and Productivity: Evidence from Teaching. *Journal of Labor Economics*, *30*(4), 749-782.

Hout, M., & Elliott, S. W. (Eds.). (2011). *Incentives and test-based accountability in education*. National Academies Press.

Hoxby, C. M. (Ed.). (2007). *The economics of school choice*. University of Chicago Press.

Ichino, A., & Maggi, G. (2000). Work Environment and Individual Background: Explaining Regional Shirking Differentials in a Large Italian Firm. *The Quarterly Journal of Economics*, 115(3), 1057-1090.

Jackson, C. K. (2013). Non-cognitive ability, test scores, and teacher quality: Evidence from 9th grade teachers in North Carolina. NBER Working Paper No. w18624.

Jacob, B.A. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, 89, 761-796.

Jacob, B.A. (2013). The Effect of Employment Protection on Teacher Effort. *Journal of Labor Economics*, 31(4), 727-761.

Jacob, B.A., & Levitt, S. D. (2003). Rotten Apples: An Investigation Of The Prevalence And Predictors Of Teacher Cheating. *The Quarterly Journal of Economics*, 118(3), 843-877.

Jacob, B.A., & Walsh, E. (2011). What's in a rating? *Economics of Education Review*, 30(3), 434-448.

Johansson, P., & Palme, M. (1996). Do economic incentives affect work absence? Empirical evidence using Swedish micro data. *Journal of Public Economics*, 59(2), 195-218.

Kane, T. J., and D.O. Staiger. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. NBER Working Paper No. w14607.

Ladd, H. F. (Ed.). (1996). *Holding schools accountable: Performance-based reform in education*. Washington, DC: Brookings Institution Press.

Luna, C., & Turner, C. L. (2001). The impact of the MCAS: Teachers talk about high-stakes testing. *The English Journal*, 91, 79-87.

Miller, R., Murnane, R., & Willett, J. (2008). Do teacher absences impact student achievement? Longitudinal evidence from one urban school district. *Educational Evaluation and Policy Analysis*, 30(2), 181–200.

Mittag, N. (2012). New methods to estimate models with large sets of fixed effects with an application to matched employer-employee data from Germany. FDZ-Methodenreport

Ose, S. O. (2005). Working conditions, compensation and absenteeism. *Journal of Health Economics*, *24*(1), 161-188.

Ost, Ben. (2014). How do teachers improve? The relative importance of specific and general human capital. *American Economic Journal: Applied Economics*, *6*(2), 127-51.

Polikoff, M. S., & Wrabel, S. L. (2013). When is 100% not 100%? The use of safe harbor to make adequate yearly progress. *Education Finance and Policy*, *8*(2), 251-270.

Reback, R., J. Rockoff, & Schwartz, H. L. (2014). Under pressure: Job security, resource allocation, and productivity in schools under NCLB. *American Economic Journal: Economic Policy*, 6(3), 207-41.

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one? *Education Finance and Policy*, *6*(1), 43-74.

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

Roza, M. (2007). Frozen Assets: Rethinking Teacher Contracts Could Free Billions for School Reform. Education Sector Reports. *Education Sector*

Stern, Nicholas (1986) *On the specification of labour supply functions.* In: Blundell, Richard and Walker, Ian, (eds.) Unemployment, Search and Labour Supply. Cambridge University Press, Cambridge, UK, pp. 143-189.

Wiswall, M. (2013). The dynamics of teacher quality. *Journal of Public Economics* 100, 61-78.

Table 1: Summary Statistics for Baseline Analytic Sample

	Full Sample	VAM Sample
	(1)	(2)
Absences	8.7	8.3
SD	[9.6]	[8.8]
Within-school SD	[9.3]	[8.2]
Within-school-year SD	[9.1]	[7.7]
Within-teacher SD	[5.1]	[4.9]
15 or more absences	0.11	0.10
2004	0.54	0.56
Failed AYP in 2003	0.46	0.48
Tested Grade	0.52	0.92
Masters	0.27	0.28
Doctorate	0.01	0.01
Selective	0.05	0.04
National Board Certified	0.10	0.12
Experience	14.1	13.9
SD	[10.0]	[9.8]
Male	0.04	0.07
White	0.81	0.81
Black	0.14	0.14
Quartile 1 VAM		0.22
Quartile 2 VAM		0.26
Quartile 3 VAM		0.26
Quartile 4 VAM		0.27
N (Teacher Years)	8,080	2,558
Schools	331	293

Notes: Teacher-years are the unit of analysis. Standard deviations [SD] are reported in brackets for non-categorical variables. The analytic samples are restricted to teachers in Title-1 schools for whom the relevant teacher- and school-level data are available. VAM = math value added measure.

Table 2: Main Difference-in-Difference Estimates of Effect of Failing AYP on Teacher Absences

Dependent Variable:	Level of Absences (A)				Log(A)	1{A > 15}		
	OLS	OLS	OLS	OLS	OLS	LPM	LPM	Logit
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Failed AYP 2003	0.25	0.24						
	(0.37)	(0.36)						
2003	Omitted							
2004	0.21	0.17	0.39	0.51	0.07	0.01	0.01	0.10
	(0.33)	(0.32)	(0.34)	(0.37)	(0.03)**	(0.01)	(0.01)	(0.12)
2004×Failed	-1.25	-1.24	-1.30	-1.37	-0.09	-0.03	-0.03	-0.35
	(0.43)***	(0.43)***	(0.44)***	(0.45)***	(0.04)**	(0.01)**	(0.02)**	(0.15)**
Adjusted R^2	0.002	0.01	0.005	0.004	0.01	0.004	0.004	
Pseudo R^2								0.01
N (teacher-years)	8,080	8,080	8,080	8,080	8,080	8,080	7,513	7,513
Schools	331	331	331	331	331	331	263	263
Teacher *X*	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School FE	No	No	Yes	Yes	Yes	Yes	Yes	Yes
School *X*	No	No	No	Yes	Yes	Yes	Yes	Yes

Notes: Standard errors are clustered by school. Teacher *X* includes a quadratic in experience and categorical indicators of race, gender, educational attainment, National Board Certification, selectivity of undergraduate institution, and grade taught. School X includes quadratics in school size and the number of full time equivalent teachers, the student-teacher ratio, and the percent of the student body that is Hispanic, black, and eligible for free or reduced price lunch. All samples are restricted to data from the 2003 and 2004 school years. Logit models are estimated using the Chamberlain (1980) conditional (FE) logit estimator, which drops units for which there is no variation in the dependent variable, and logit coefficients are reported. 1{.} is the indicator function. LPM = Linear Probability Model. *** p<0.01, ** p<0.05, * p<0.1.

Table 3: Robustness Checks of Effect of Failing AYP on Teacher Absences

Specification:	Lagged Absences	2-way Teacher-School FE	Balanced Panel	Balanced Panel + Teacher FE	< 50 Abs. Sample Restriction	Discontinuity Sample Restriction	RD Estimate (2SLS)	Quadratic School Time Trends	Event Study
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1998×Failed									0.23
									(0.38)
1999×Failed									-0.28
									(0.42)
2000×Failed									-0.08
									(0.46)
2001×Failed									-0.11
									(0.43)
2002×Failed									0.15
									(0.44)
2003×Failed									0.29
									(0.46)
2004×Failed	-1.10	-1.09	-1.40	-1.06	-0.98	-1.09		-1.65	-1.07
	(0.53)**	(0.86)	(0.47)***	(0.47)**	(0.34)***	(0.45)**		(0.65)**	(0.48)**
Failed							-1.09		
							(0.97)		
Adjusted R^2	0.02	0.15	0.004	0.01	0.004	0.002	0.01	0.03	0.01
N	6,149	8,080	7,898	7,898	7,991	6,396	4,187	28,456	28,456
Schools	313	331	331	331	331	251	294	367	367
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Teacher FE	No	Yes	No	Yes	No	No	No	No	No

Notes: Standard errors are clustered by school. Control variables include a quadratic in experience and categorical indicators of race, gender, educational attainment, National Board Certification, selectivity of undergraduate institution, grade taught, quadratics in school size and the number of full time equivalent teachers, the student-teacher ratio, and the percent of the student body that is Hispanic, black, and eligible for free or reduced price lunch. The samples in columns 1-6 are restricted to data from the 2003 and 2004 school years. The sample in column 7 uses only 2004 data while the sample in columns 8 and 9 uses data from 1997-2004 and includes a full set of year FE. The balanced panels in columns 4 and 5 exclude teachers who changed schools between 2003 and 2004. RD = Regression Discontinuity. The "discontinuity" sample in column 6 is restricted to schools whose distance to the AYP cutoff was between the 10^{th} and 90^{th} percentiles of the distribution. Column 7 applies a "simple" fuzzy RD 2SLS estimator (Angrist & Pischke, 2009, p. 261). *** p<0.01, ** p<0.05, * p<0.1.

Table 4: Heterogeneity in Effects of Failing AYP on Teacher Absences

	Baseline Sample		VAM Sample		
	(1)	(2)	(3)	(4)	(5)
2004	1.19		0.34	-0.81	
	(0.70)*		(0.62)	(1.10)	
2004×Failed	-1.71		-0.82	1.32	
	(0.93)*		(0.72)	(1.53)	
2004× Failed×Tested	1.12	1.62			
	(0.91)	(1.06)			
2004×Failed×Masters	0.14	0.24			
	(1.00)	(1.05)			
2004×Failed×Doctorate	1.36	3.84			
	(8.84)	(7.82)			
2004×Failed×Selective	-3.92	-4.13			
	(2.16)*	(2.17)*			
2004×Failed×NBC	-0.27	-0.05			
	(1.20)	(1.33)			
2004×Failed×Experience	-0.01	-0.02			
	(0.04)	(0.05)			
2004×Failed×Male	1.28	1.34			
	(1.27)	(1.42)			
2004×Failed×VAM Q2				-2.05	-4.10
				(2.01)	(2.91)
2004×Failed×VAM Q3				-3.64	-6.69
				(2.07)*	(3.31)**
2004×Failed×VAM Q4				-2.73	-5.74
				(1.95)	(3.64)
School FE	Yes	No	Yes	Yes	No
School-by-Year FE	No	Yes	No	No	Yes
Adjusted R^2	0.004	0.004	0.01	0.01	0.01
N (teacher-years)	8,080	8,080	2,558	2,558	2,558
Schools	331	331	293	293	293

Notes: Standard errors are clustered by school. All models condition on *X* and the appropriate sub interactions (e.g., 2004×*X*, Failed×*X*). The models are estimated using data from the 2003 and 2004 school years. *** p<0.01, ** p<0.05, * p<0.1.