NBER WORKING PAPER SERIES

SCHOOL ACCOUNTABILITY, LONG-RUN CRIMINAL ACTIVITY, AND SELF-SUFFICIENCY

Ozkan Eren
David N. Figlio
Naci H. Mocan
Orgul Ozturk

Working Paper 31556
http://www.nber.org/papers/w31556

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2023

School Accountability, Long-Run Criminal Activity, and Self-Sufficiency
Ozkan Eren, David N. Figlio, Naci H. Mocan, and Orgul Ozturk
NBER Working Paper No. 31556
August 2023
JEL No. H0,I0

ABSTRACT

This paper examines the impact of school accountability on adult crime and economic self-sufficiency. We employ a unique source of linked administrative data from a Southern state and exploit exogenous variation generated by the state's accountability regime. Our findings indicate that a school's receipt of a lower accountability rating, at the bottom end of the ratings distribution, decreases adult criminal involvement. Accountability pressures also reduce the propensity of students' reliance on social welfare programs in adulthood and these effects persist at least until when individuals reach their early 30s. Further examination reveals that our results are consistent with an explanation related to improvements in human capital accumulation.

Ozkan Eren
Department of Economics
University of California, Riverside
Sproul Hall
Riverside, CA 92521
ozkan.eren@ucr.edu

Naci H. Mocan
Department of Economics
Louisiana State University
2439 BEC
Baton Rouge, LA 70803-6306
and NBER
mocan@lsu.edu

David N. Figlio
University of Rochester
220 Hutchinson Road
Rochester, NY 14611
and NBER
david.figlio@rochester.edu

Orgul Ozturk
University of South Carolina
1014 Greene Street
Columbia SC 29208
odozturk@moore.sc.edu

1 Introduction

School accountability systems evaluate schools on the basis of aggregate student performance measures. These systems generate rewards and sanctions under the premise that various combinations of carrots and sticks can improve the focus and productivity of public schools.[1] There is evidence that school accountability systems have some desired outcomes: numerous studies find large gains of test-based accountability on student test scores (Ladd 1999; Carnoy and Loeb 2002; Hanushek and Raymond 2004; Figlio and Rouse 2006; Chiang 2009; Rockoff and Turner 2010; Dee and Jacob 2011; Rouse et al. 2013; Reback et al. 2014 for US evidence, and Nunes et al. 2015; Andrabi et al. 2017; Cilliers et al. 2021 for international evidence). These accountability ratings have effects that go well beyond the school system; for instance, Figlio and Lucas (2004) show that school accountability ratings affect housing markets.

But, of course, it may be that school accountability systems only improve performance on the metrics and domains for which schools are being held accountable. There is ample evidence that, when faced with expectations of boosting test performance, schools respond by focusing on particular subjects and certain groups of students most central to accountability ratings, and by manipulating the pool of test-taking students (Cullen and Reback 2006; Figlio 2006; Figlio and Getzler 2006; Reback 2008; Neal and Schanzenbach 2010), and artificially inflating measured test performance (Figlio and Winicki 2005).

For this reason, and also the potential that short-term effects do not necessarily predict long-term effects of policies even absent manipulative behavior, it is important to observe whether school accountability measures have long-term benefits for students, or if the observed benefits regarding test scores are merely transitory. To date, however, there exists very little evidence on longer-run effects of school accountability, largely due to the paucity of data linking childhood education to outcomes in adulthood. We are aware of only one economics paper studying longer-run effects of school accountability policies

[1] In the US, the 2002 No Child Left Behind Act (NCLB) mandated that all US states introduce some form of test-based school accountability, continuing a trend of state accountability policies that began in the 1990s. The Every Student Succeeds Act (ESSA), which was signed into law in 2015, replaced the NCLB. Under the ESSA, states have more responsibility over their accountability systems and standards.

(Deming et al. 2016), which investigates the role of accountability on educational attainment and early-career (through age 25) labor market outcomes. The incentives investigated by Deming et al. (2016) in the Texas context are particularly salient for schools on the margin of high accountability ratings and stand in contrast to more recent accountability regimes under the No Child Left Behind.[2] Therefore, we know quite little about the potential long-term consequences of school accountability for schools on the margin of low accountability ratings, the margin that has been most often studied with regard to short-run outcomes, and the set of schools that educate larger fractions of the most vulnerable students.

We use a unique source of linked administrative data from South Carolina to investigate the effects of school accountability on adult crime and economic self-sufficiency (measured by reliance on social welfare programs), two outcomes that are particularly important for the population of students attending schools on the margin of low accountability ratings. We are able to study these outcomes deeper into adulthood, as late as age 34. South Carolina's accountability system, introduced in 2000, permits quasi-experimental identification using regression discontinuity (RD) design and local randomization approach**.**

As part of the accountability system implemented by the South Carolina Department of Education (SCDE), all public schools are evaluated according to a set of continuous performance metrics, which are then converted into discrete school ratings (e.g., Unsatisfactory, Average, Excellent) based on sharp cutoffs. This information is made public, published in at least one daily newspaper of general circulation in the area, and school report cards are mailed to parents soon after the release. The SCDE uses these performance ratings to both reward and sanction the schools. High ratings are associated with additional funding, while schools that receive low ratings face serious consequences such as leadership change, restructuring, and state takeover.

[2] Deming et al. (2016) find that students in schools that were at risk of receiving a low performing rating were more likely to attend and graduate from a four-year college and had higher earnings at age 25. In contrast, low achieving students in schools that were close to receiving a high rating had significantly lower postsecondary attainment and earnings. The authors attribute these conflicting results to schools' heterogeneous responses to the incentives created by accountability regime in Texas in the 1990s. More precisely, schools facing pressure to achieve a high rating typically had a smaller fraction of low performing students and thus were more likely to classify these students into test-exempt special education categories. Given the size of low performing students, such strategic responses were less prevalent in schools at risk of receiving a low performing rating.

We find evidence that the identification assumptions necessary for the RD framework to be credible are met in the South Carolina accountability context. We provide several robustness analyses and validity tests supporting these identifying assumptions throughout the paper. We further complement our analysis using local randomization approach which changes the parameter of interest from the RD treatment effect at the cutoff to the RD treatment effect in the neighborhood around the cutoff where local randomization is assumed to hold. This alternative method safeguards against potential complications related to estimation and inference that may arise in standard RD analysis when the running variable is discrete and only a few mass points are present (i.e., many observations sharing the same values of the running variable). To the best of our knowledge, this is one of the first empirical applications of the local randomization approach. Finally, we present several regression results separately for female and male students. Our analysis of effects by gender is motivated by prior research which shows that girls and boys may respond differently to social programs, family conditions and school environment (Bertrand and Pan 2013; Garcia et al. 2018; Autor et al. 2019; Bald et al. 2022).

The results indicate that students who attended high schools, which had lower accountability ratings at the bottom end of the ratings distribution, are less likely to engage in criminal activity in adulthood and are more likely to be economically self-sufficient. Specifically, these students are 1.8 percentage points less likely to have ever been arrested in adulthood (an 8 percent reduction relative to the control mean) and are 2.8 percentage points less likely to rely on social welfare programs in adulthood between the ages of 18 and 34 (a 4.5 percent reduction relative to the control mean). The discontinuity estimates for both adult crime and the receipt of government assistance are more pronounced and precisely estimated for female students. The estimated effect of a school receiving a lower accountability rating on long-run outcomes of the school's students is small and statistically indistinguishable from zero at the top end of the ratings distribution.

We find little evidence that the South Carolina accountability system affected either endogenous mobility of students or strategic responses of schools. We do, however, observe that while graduation

rates did not change appreciably as a consequence of accountability pressure in the South Carolina context, schools' academic standards and student performance improved. These changes took place without significant adjustments in teacher quality, teacher turnover, per pupil spending or the replacement of school principals. It appears, therefore, that South Carolina's accountability system led to lasting improvements in the life outcomes of students attending schools at risk of poor accountability ratings.

2 The South Carolina Accountability System

As part of South Carolina's accountability system, launched in 2000, all public schools are assigned one of five performance categories: (i) Unsatisfactory, (ii) Below Average, (iii) Average, (iv) Good, and (v) Excellent. These performance categories are based on a continuous index, known as accountability performance score. During the period we analyze a high school's score is calculated using the weighted sum of four components: the percentage of tenth grade students who meet the standards of exit examination (20 percent of the overall score), longitudinal exit exam performance (30 percent), the percentage of students eligible for merit-based (LIFE) scholarship to a four-year institution (20 percent), and the graduation rate (30 percent).[3] A school's accountability score (ranging between 1 and 5) determines that school's performance category. For example, schools earning fewer than 2.2 points received an Unsatisfactory rating, while schools with 2.2 points received a Below Average rating.[4]

Several aspects of the rating formula were revised in the early years of the accountability system. For example, the last component (graduation rate) was added to the calculation of the overall score beginning with the 2002-2003 academic year.[5] There were also other changes and revisions to the calculation of the

[3] Between 1986 and 2005, the state administered the Basic Skills Assessment Program, which is a minimum competency exam, as its exit exam. Students had to pass all three subjects (reading, writing and math) to meet the exit exam standards. Longitudinal exit exam performance of schools was determined by the fraction of students who passed the exit exam by the spring of twelfth grade. The eligibility for LIFE scholarship was based on the fraction of students meeting both the GPA and SAT/ACT criteria established by the state.

[4] Over the period from 2000 to 2002, schools were rated as Excellent for an overall performance score of 3.4 and above, Good for scores between 3.0 and 3.3, Average for scores between 2.6 and 2.9, Below Average for scores between 2.2 and 2.5 and Unsatisfactory for scores below 2.2.

[5] A high school's overall performance in the first two years of the accountability system was calculated using the percentage of tenth grade students who meet the standards of exit examination (30 percent), longitudinal exit exam performance (30 percent) and the percentage of students eligible for LIFE scholarship to a four-year institution (40 percent).

accountability performance score right after the first year of its implementation. These revisions included changes in the eligibility criteria for LIFE scholarship, and regarding the status of students taking the exit exam in grades other than the tenth grade.

These changes to how the scores are calculated over the first few years of the program are important because they made it difficult for schools to manipulate their scores around the rating cutoffs. In addition, South Carolina's accountability system limited the room for schools' strategic behavior by allowing exclusion of students from high-stakes testing only if students' circumstances related to disabilities and Limited English Proficiency were in accordance with federal guidelines (South Carolina Education Oversight Committee 2000-2003). This stands in sharp contrast to earlier accountability regimes such as that analyzed in Deming et al. (2016).

The SCDE tied the performance ratings to both rewards and sanctions. Those schools that received a Below Average or Unsatisfactory rating are required to develop improvement plans with the assistance of an external review team, members of which comprised representatives from SCDE and selected school districts, retired educators, parents, and other community members. Schools' improvement plans must focus on strategies that aim to increase academic performance, offer professional development activities for teachers, and include a timeline for progress. Upon recommendation of the review team, the state (and the district) can also assign teacher/principal specialists to schools designated as Below Average or Unsatisfactory. These education specialists provide different forms of assistance, including developing research-based instructional strategies targeting specific needs of students in the school, leading teacher development groups, providing support in the form of observation with feedback, and modeling. The SCDE also established grant programs for improvement in schools which are rated as Unsatisfactory and Below Average. On the reward side, high ratings are associated with additional funding in which the maximum amount of money a school can receive is equivalent to a school's per-pupil allocation.[6] These funds are generally used for professional development purposes.

[6] For example, these payments totaled around $1 million in the 2002-2003 academic year.

The SCDE releases key information from school report cards to the parents and general public no later than mid-November which is roughly two weeks after the distribution of report cards to schools. The accountability system is expected to create pressure for school administrators to improve student achievement. Such pressures may stem from a variety of sources, ranging from intensive scrutiny and supervision to social stigma, from threat of job loss to disutility resulting from failure to fully foster the development of children.

As described above, in terms of targeted assistance, the accountability system in South Carolina treats all low performing schools similarly. Because of its consequences, however, accountability pressures are expected to be stronger for schools rated as Unsatisfactory. For example, the SCDE made it clear that schools receiving an Unsatisfactory rating, absent of adequate progress, are susceptible to leadership change, restructuring, and state takeover (South Carolina Education Oversight Committee 2000-2003, Article 15). Along these lines, a growing number of studies document that accountability pressures placed on schools are much stronger at the bottom end of the ratings distribution than at the top (Rockoff and Turner 2010; Rouse et al. 2013; Dizon-Ross 2020).

3 Data

The data for this study are compiled from several different sources. The first one is administrative records from the South Carolina Department of Education. The data include student race, gender, free/reduced lunch status and age, and test score information from selected grades. Unique identification numbers allow us to track all students through their tenure in the public school system from the fall of 2000 onwards.

Our main crime data come from the South Carolina State Law Enforcement Division (SLED) and include the universe of detailed arrest records from 2000 to 2017. For each offender file, we have basic demographic information on the arrestees, offense date, and the type of crime committed. We complement these data with conviction records that resulted in incarceration which are obtained from the South

Carolina Department of Corrections over the same period. We also utilize the administrative records from the South Carolina Department of Social Services (SCDSS), available from 2000 to 2019, to gather information on enrollment in social welfare programs. Using unique identification numbers, we are able to link individuals' records in all these four data sets. Finally, we rely on publicly available school report cards for data on schools' performance ratings (Unsatisfactory, Below Average, etc.), their overall score which determines the performance rating, the components of the overall score, and several other school level attributes, such as measures of disciplinary climate, teacher turnover, and so on.

Our sample consists of first-time ninth graders from the 2000-2001 to 2002-2003 academic years, roughly corresponding to the cohorts born between 1985 and 1988. These cohorts were "treated" by the accountability system during its early years of implementation. We thus aim to minimize confounders that may arise from potential adjustments that could have been made by schools to manipulate their performance scores around the rating cutoffs. As discussed in Section 2, the specifics of the formula which generates the accountability points were revised repeatedly in the first few years of the accountability system's adoption. This created a somewhat moving target for the schools, and therefore made it difficult for them to strategically adjust their behavior at the margins of the rating cutoffs. We assign students to the first high school they attended. Doing so circumvents concerns related to endogenous responses from students and parents such as transferring to another school following a low performance rating and it gives our results an intent-to-treat interpretation.

One of our main outcomes of interest is an indicator for whether the individual was ever arrested as an adult, which we can observe up to age 32. We employ a similar indicator for adult incarceration. We have access to complete administrative records from the South Carolina Department of Juvenile Justice beginning in 2003. The upper age of juvenile court jurisdiction over our analysis period was 16 and an overwhelming majority of students were ages 14 to 15 by the time they entered high school. As a result, for these cohorts, we cannot analyze the impact of school accountability on juvenile crime. Records from the SCDSS allow us to construct two measures of economic self-sufficiency: whether the student ever

received food stamps as an adult (renamed Supplemental Nutrition Assistance Program (SNAP) in 2008) and whether the student ever received Temporary Assistance for Needy Families (TANF) as an adult. Food stamps program has a significantly larger base of participation than TANF and provides a steady stream of benefits to households that are income and asset-eligible, as well as able-bodied adults without dependents.[7] Given that SCDSS is available through 2019, we can observe reliance on social welfare programs up to age 34.

Table 1 presents the descriptive statistics for a total of more than 160,000 students from 194 unique high schools. We show tabulations for the full sample, as well as by schools' performance ratings. As displayed in Panel A, black and white students comprise 41 and 55 percent of all students, respectively and the percentage of black students is decreasing along accountability ratings. Similarly, there is a negative relationship between the fraction of free-lunch eligible students and schools' ratings. The opposite pattern is displayed between the fraction of students who were proficient in eight grade subject tests. The eighth grade standardized test scores were missing for an overwhelming majority of the analysis sample. As a result, we use discrete achievement indicators (e.g., proficient; advanced), which are available for more than 70 percent of the analysis sample, to proxy for subject-specific eighth grade achievement level in math and English Language Arts (ELA).

As shown in Panel B, 24 percent of students, who attended a high school that was rated as Unsatisfactory, were arrested as an adult. About 8 percent were incarcerated (Column 2). The gap between the arrest and incarceration rates of students in schools rated as Unsatisfactory and Average is slightly more than 2 percentage points (Columns 2 and 4). Driving under influence, disorderly conduct, possession of drugs and shoplifting are the most common types of arrests in the data.

The first column of Table 1 reveals that 51 (12) percent of students in our sample used food stamps (TANF) as an adult. Not surprisingly, consistent with the primary target populations of these programs,

[7] The total cost of the food stamps program was around 60 billion dollars in 2019 (U.S. Department of Agriculture 2019). The states spent about 31 billion dollars in federal and state funds under the TANF (U.S. Department of Health and Human Services 2019).

the reliance was mostly prevalent among female students. Fifty-four percent of those who ever received a food stamp between the ages of 18 and 34 are female (i.e., 54 % of 0.513 in Column 1 are female), and 81% of those who were a recipient of TANF at least once between the ages of 18 and 34 are female (81% of 0.122 are female). The fraction of individuals receiving government assistance are disproportionately associated with low performing schools. For example, while 37 percent of individuals who attended schools with an "Excellent" rating received food stamps as an adult, the rate is about 74 percent among individuals who attended "Unsatisfactory" schools. Panel C of Table 1 reveals that, compared to high rated schools, schools at the bottom end of the ratings distribution had higher teacher turnover and lower teacher quality (measured by the fraction of teachers with an advanced degree). Per pupil spending in these schools was higher.

Finally, Figure 1 displays the relationship between schools' accountability ratings and their overall performance scores. It is evident that the rating cutoffs were strictly enforced over our sample period.

4 Empirical Methodology

4.1 Regression Discontinuity and Local Randomization Approach

To evaluate the effects of receiving a lower accountability rating on long-run outcomes of individuals who were students in these schools, we leverage the discontinuous relationship between accountability ratings and performance scores that determines the ratings (as depicted in Figure 1) and estimate the following equation

$$Y_{ijc} = \beta_0 + \beta_r A^r_{jc} + \lambda f(S_{jc}) + \gamma X_{ijc} + \epsilon_{ijc} \tag{1}$$

where Y_{ijc} is the outcome of interest such as an indicator that takes the value of one if the student ever used food stamps as an adult between the ages of 18 and 34 (i denotes the student, j the school, and c the high school entry cohort). A^r_{jc} is an indicator for the accountability rating received by the school (r denotes the rating). $f(S_{jc})$ is a quartic in overall accountability score. X_{ijc} is a vector of observed

covariates (indicators for gender, race and free/reduced price lunch, age student was first found in public school, cohort fixed effects, the percentage of ninth grade students who were female, black, free/reduced lunch eligible and average age first found in public school) and ϵ_{ijc} is the error term. The control function $f(\cdot)$ is also interacted with cohort fixed effects to capture the changes in the calculation of the overall performance score, implemented by the state, over the sample period. Standard errors, clustered at the school level, are reported throughout the analysis.

Because of the policy relevance and owing to growing evidence on the relationship between incentives and accountability pressures (Rockoff and Turner 2010; Rouse et al. 2013; Dizon-Ross 2020), we concentrate on students at the bottom end of the school ratings distribution throughout the paper although we also present the main results for students at schools where accountability pressures were much weaker, i.e., at the top end of the ratings distribution (Section 5.2). To improve efficiency, we estimate the impact of receiving a lower accountability rating by pooling schools from the bottom three groups together (Unsatisfactory, Below Average, and Average). More precisely, we take all schools in the middle group (those rated as Below Average) and divide them into two groups based on whether their overall accountability score places them below or above the median for that rating in a given year. We then assign above (below) median schools as a comparison group for those rated as Average (Unsatisfactory). As a result, A^R_{jc} in equation (1) becomes a simple indicator function denoting a lower accountability rating assigned to schools that are in the bottom three groups. The RD estimates from such grouping exercise represents a weighted average of the effects at two individual cutoffs and is local to schools in the close vicinity of the rating cutoffs (Rockoff and Turner 2010; Dizon-Ross 2020). We also present the results obtained from analyzing the impact of accountability ratings at each separate cutoff in Section 5.3. This alternative modeling, which allows an explicit comparison between Unsatisfactory and Below Average schools, arguably nets out any potential effect of targeted assistance to schools because, as noted in Section 2, all schools rated as Unsatisfactory or Below Average received such assistance.

Cattaneo et al. (2018) caution against using local polynomial methods when the running variable is

discrete and only a few mass points are present (i.e., many observations sharing the same values of the running variable). This is because extrapolation towards the cutoff performs poorly in the presence of few mass points. Put differently, the discrete nature of the running variable in conjunction with small number of mass points makes local polynomial methods difficult to justify. We address this issue in two ways. First, as noted, we rely on a global polynomial fit as the preferred specification and check the sensitivity of the results using varying orders of polynomials (and local linear regression).

Second, we employ the local randomization approach, which is based on the assumption that placement above or below the cutoff within a very close window of the discontinuity is as good as random (i.e., treatment assignment can be assumed experimental). This method effectively changes the parameter of interest from the RD treatment effect at the cutoff to the RD treatment effect in the neighborhood around the cutoff where local randomization is assumed to hold (Cattaneo et al. 2018). Importantly, given the discrete nature of the running variable, the local randomization approach does not necessarily require a window selection procedure because the location of the minimum window is known — it is the interval of the running variable that contains two consecutive mass points where the treatment changes from zero to one. To the extent that local randomization holds, it also must be true for the minimum window which is the rationale for the focus on these two consecutive mass points (without conditioning on the running variable). The local randomization approach has appealing features in discrete settings and estimation and inference based on standard experimental methods are applicable. However, the small number of observations inside the window is likely to limit its statistical power. Consequently, we use this alternative framework as a robustness check for our main findings (Section 5.3).

4.2 Validity of the Regression Discontinuity Research Design

In our framework, the key identifying RD assumption is that, conditional on a flexible control for overall accountability score, the assignment of a school rating is exogenous. This assumption, although inherently untestable, does yield testable implications. First, we would expect pre-determined individual charac-

teristics to be smooth through the cutoffs. Panel A of Table 2 reports the estimated discontinuities in baseline covariates for the full sample. The coefficient estimates are all small in magnitude and none is statistically different from zero. These pre-determined characteristics continue to be balanced separately for female and male students (Panels B and C). Appendix Table A1 tests similar discontinuities using several school-level measures and shows that observable school characteristics are also balanced around the cutoffs.[8]

Second, the density of schools should be continuous. We formally test the smoothness of the density and fail to reject the null hypothesis of a continuous distribution (p-value=0.34).[9] Figure A1 displays the distribution of the overall accountability scores for schools at the bottom end of the ratings distribution. These results lend support to the assumption that, after controlling for the accountability score in the specifications, whether a school received a high or low rating is as good as random. To further minimize concerns on manipulation, we present the estimated effects by employing a donut-RD (Section 5.2).

As a preliminary step, we provide a graphical representation of discontinuities at each separate cutoff at the bottom end of the ratings distribution. The graphs of raw outcomes (adult arrest and participation in food stamps/TANF), displayed in Figure 2, show non-trivial differences in average long-run outcomes and trends. Figure 3 plots the residuals from a regression of adult arrest (Panel A) and participation in social welfare programs (Panel B) on a quartic polynomial in overall accountability performance score (interacted with cohort fixed effects). Fitted values from a locally weighted polynomial regression are superimposed over these residuals. Appendix Figure A2 displays these residuals by student's gender. There are visible discontinuities in both outcomes at the Unsatisfactory-Below Average and Below Average-Average rating cutoffs.

Finally, it bears noting that we can observe all criminal activity resulting in arrests, and dependence on social welfare programs insofar as students did not leave the state. Differential attrition on either side

[8] These regressions are weighted by the number of observations that underlie each school-by-cohort cell.

[9] Given the discrete nature of the running variable, we test for manipulation by employing the test proposed in Frandsen (2017) and use the Stata package rddisttestk.

of the rating cutoffs would invalidate our identification, but such sample selection is unlikely to be an issue. In fact, using the American Community Survey data, we find that less than 10 percent of the adult population born in South Carolina between 1985 and 1988 left the state in early adulthood (age 30 or younger).

5 Results

5.1 Baseline Results

We present our baseline results on the relationship between lower accountability ratings and adult crime in Table 3. Column 1 reports the impact by controlling for only cohort fixed effects. Column 2 shows the results when student characteristics are included. Finally, Column 3 presents the results by further adding grade level school characteristics. Columns 1-3 reveal that the RD estimates of the effect of receiving a lower accountability rating on long-run criminal activity is not sensitive to the inclusion of any control variables, providing assurance as to the credibility of the identification strategy.

Focusing on our preferred specification in Column 3, we find that lower accountability rating of a school decreases the likelihood of its student ever being arrested as an adult. Specifically, students in schools that were located just below the rating cutoff are 1.8 percentage points less likely to be arrested in adulthood in comparison to students who attended schools that were just above the cutoff. This represents a decrease of 8 percent relative to the control mean. Columns 4 and 5 of Table 3 report the results by student's gender. The discontinuity estimates are similar in magnitude for male and female students, but the impact for female students is twice the size of that for male students (13 and 6 percent for female and male students, respectively) when the coefficients are benchmarked relative to gender-specific control means.

We also examine the effect of receiving a lower accountability rating on students' likelihood of being incarcerated. The point estimate, reported in the last column of Table 3, is small in magnitude and

statistically insignificant. Further examination of arrests by types of crimes reveal that the discontinuity estimates observed in Table 3 are largely driven by alcohol and drug related crimes.[10] As noted, arrests related to alcohol and possession of drugs constitute the most common types of offenses in our data. Considering that such offenders make up less than 4 percent of the prison population in the U.S. (Carson 2014), a null finding for incarceration is unsurprising.

Table 4 displays the results of the analyses where we investigate the effect of receiving a lower accountability rating, at the bottom end of the distribution, on students' economic self-sufficiency in early adulthood. The results are presented for the full sample, as well as by gender. Similar to those in Table 3, the point estimates in Panel A are all negative across columns, but we find a large and statistically significant coefficient estimate only for female students. Lower accountability ratings decrease the propensity to rely on social welfare programs for female students in adulthood by 4.2 percentage points, which represents a 6 percent decrease relative to the control mean. Panels B and C present the same set of results separately for the receipt of food stamps and TANF, respectively. The effect of a lower school rating on the use of food stamps for female students is significant (Column 2). The food stamps benefit has been an important source of income for recipients in South Carolina where the average monthly SNAP benefit is roughly equivalent to one-fourth of the total gross income recipients reported over the period 2010-2019 (SNAP Quality Control Files, Mathematica Policy Research, Inc.). The coefficient estimate for TANF participation is not statistically different from zero (Panel C).[11]

To investigate if the results are driven by adults who are younger (25 years old or younger), or older (between 26 and 31-34, depending on the outcome), we estimated the models within these two age groups. The results, reported in Appendix Table A2, reveal that a school's receipt of a lower accountability rating leads to decrease in the propensity for adult crime both in the age group of 18-25, and also when the

[10] The discontinuity estimates are as follows: -0.013 (s.e.=0.007) for alcohol and drug related crimes, -0.002 (s.e.=0.004) for property crimes and -0.004 (s.e.=0.003) for violent crimes.

[11] South Carolina has a full SNAP ban in place since 1996 for offenders convicted of certain felony crimes, indicating that those with an arrest record are less likely to receive future SNAP support. Nevertheless, we created a binary variable that takes the value of one if the student participated in welfare programs and also got arrested in adulthood. The discontinuity estimates from this exercise are -0.021 (s.e.=0.007) and -0.008 (s.e.=0.012) for female and male students, respectively.

individuals are older than 25. In both cases the estimated coefficients are negative but imprecisely estimated.

Appendix Table A2 also shows that for females, being affiliated with a school that has received a lower rating, in comparison to otherwise similar schools which received a higher rating, has a negative impact on welfare participation during young adulthood (18-25), as well as when older than 25. Figure 4 shows that this impact on welfare receipt exists at any age, albeit being insignificant at ages 23 and below.[12] More specifically, Panels B and D of Figure 4 show that having attended a lower-rated school (which was exposed to accountability pressures) has a negative impact on the probability of being the recipient of welfare assistance in adulthood for females, with more pronounced effects between the ages of 25 and early 30s.

To put these estimates in perspective, we compare our estimates to other studies in the related literature. For example, Billings et al. (2014) find that a 10 percentage point increase in the share of minorities in a student's assigned middle school increases adult arrest rates by 7 percent. The impact of receiving a lower rating we identify here is about the same size. Currie et al. (2001) show that a one percentage point decline in the unemployment rate accounted for about 10 percent of the decrease in food stamps participation over the period from 1993 to 1998. Our estimated effect of school accountability on the receipt of social assistance for female students is slightly above half of the effect resulting from a one percentage point decline in unemployment rate reported by Currie et al. (2001). Similarly, the RD estimate for adult crime (use of food stamps) corresponds to 25 (12) percent of the raw gap in these outcomes between the schools rated as Unsatisfactory and Excellent (Table 1).

5.2 Robustness Checks

We conducted a number of sensitivity checks to examine the robustness of the results. These results are reported in Appendix Table A3. First, we re-estimated Equation (1) using both quadratic and cubic

[12] Each point in each panel comes from a separate regression where the dependent variable takes the value one if the individual enrolled in social programs by the given age.

specifications for $f(S_{jc})$, as well as by limiting the sample to schools for which the performance scores were within specific distances from the Unsatisfactory and Below Average cutoffs (Columns 1-3). As noted above, when the number of mass points is very small, local polynomial methods are difficult to justify (Cattaneo et al. 2018). Nevertheless, we also estimated the effect of receiving a lower accountability rating using a local linear regression (Column 4). The discontinuity estimates are similar to our baseline results. Second, we controlled for eighth grade subject-specific standardized test achievement indicators, i.e., indicators for whether the student was labeled proficient in math and ELA (Column 5).

Third, recall that we limit our analysis to ninth graders from the 2000-2001 to the 2002-2003 academic years. Such restriction arguably minimizes concerns related to strategic responses of schools because the formula generating the performance scores were revised repeatedly in the early years of the accountability regime (Section 2). In later years the evaluation criteria remained stable, giving opportunity for schools to adjust their behavior strategically. However, extending the data to include more recent ninth grade cohorts (2003-2004 to 2005-2006) do not change the results in a meaningful way (Column 6). Appendix Table A4 reports the estimated discontinuities in baseline covariates when adding these additional cohorts. Despite a more than twofold increase in sample size, statistical inference on covariate balance tests remains intact. Relatedly, we focused on only the first cohort — the ninth graders who started high school in the 2000-01 academic year, in which schools had no opportunity to respond to whatever ratings they would have received. This sub-sample generated the same inference.[13]

Fourth, we re-run our baseline specification by employing a donut-RD where we remove schools that received 2.2 and 2.6 points (thresholds for Below Average and Average ratings, respectively) over the sample period. The estimated effects reveal that the results are not sensitive to dropping observations at the points of discontinuity (Column 7). Fifth, we collapsed the data at the school-by-cohort level and estimated the impacts of receiving a lower rating. These aggregate level regressions are also weighted by the number of students that underlie each school-by-cohort cell (Column 8). The discontinuity estimates

[13] The point estimates are -0.027 (s.e.=0.012) and -0.043 (s.e.=0.020) for adult crime (using 18,945 students from the first cohort) and welfare receipt (using 9,035 female students from the first cohort), respectively.

from these alternative specifications are very similar to those reported in Tables 3 and 4. Sixth, we performed a placebo test where we assigned schools their four year-ahead accountability ratings and performance scores.[14] As shown in the last column of Appendix Table A3, the point estimates from this exercise carry opposite signs and they are statistically insignificant.

We conducted another placebo exercise in which we took all students and the schools they are affiliated with, and randomly changed the "treatment status." More specifically, we took the actual values of schools' treatment along with accountability scores for a given year and re-distributed them randomly across schools. After this random assignment, some students who attended lower-rated schools would be considered as if they attended about-the-cutoff schools, and vice versa. We repeated this process 1,000 times, running our models after each random re-allocation. The estimates obtained from this exercise are distributed around zero. Figure 5 displays these distributions along with the estimates obtained from our models that use the true school rating assignments (represented by the vertical lines). We report the percentage of placebo estimates that are smaller than the baseline effects on the x-axis. In all cases, the location of the true estimates indicates that the likelihood of finding an effect merely by chance is unlikely.

We also explore the potential heterogeneity in the effects of receiving a lower accountability rating by student's proficiency level in eighth grade standardized tests. In absolute value, the estimated impact for adult crime (receipt of government assistance) is larger (smaller) in magnitude for students who were labeled proficient in either of these subjects (Appendix Table A5).[15]

Finally, we investigate the relationship between accountability and long-run outcomes at the top end of the ratings distribution (Excellent, Good, and Average).[16] Appendix Table A6 reports the estimated

[14] Ideally, we would like to use pre-accountability information for a falsification exercise, however, we do not have such pure pre-accountability data. Note also that each school-by-year observation can be matched to their future ratings only if they stay at the bottom of the ratings distribution at $(t+4)$.

[15] The lack of precision of these estimates is likely the result of smaller sample sizes because eighth grade subject-specific achievement indicators are missing for about 30% of our sample.

[16] In this specification, we take all schools in the middle group (rated as Good) and divide them into two groups based on whether their overall accountability score places them below or above the median for that rating in a given year. We then assign above (below) median schools as a comparison group for those rated as Excellent (Average).

discontinuities in baseline covariates. There is some evidence against covariate balance at the ratings cutoff and thus these results should be interpreted with caution. With this caveat in mind, we do not find any large and significant impact of school rating on long-run outcomes in this range (Appendix Table A7). Our conclusions are not altered in a meaningful way when we estimate the effects of accountability pressures by student's eighth grade proficiency level at the top end of the ratings distribution.

5.3 Heterogeneous Effects by Performance Ratings and Local Randomization Results

In this section, we analyzed the impact of accountability ratings at each separate cutoffs at the bottom end of the ratings distribution (Table 5). The coefficient estimates on long-run outcomes are negative for students in schools that received an Unsatisfactory or Below Average rating although these coefficients are less precisely estimated than those in models in which schools from the bottom thresholds are pooled together. As further shown in the table, the effects of accountability pressures also appear to be more pronounced for students at schools that were rated as Unsatisfactory. Specifically, students in schools rated as Unsatisfactory were 6.6 percentage points less likely to be ever arrested as an adult than students from schools rated as Below Average (-0.095 vs. -0.029). A test of equality between these two coefficients is rejected (p-value=0.00). Using the coefficients reported in Column 3, a similar comparison for reliance on social welfare programs implies a reduction of 3.6 percentage points for female students in schools rated as Unsatisfactory (-0.077 vs. -0.041).

To further check the sensitivity of our results, we employ the local randomization approach. As discussed in Section 4.1, when only a few mass points of the running variable are present bandwidth selection (and therefore local polynomial models) makes little sense. Cattaneo et al. (2018) propose estimating treatment effects in the interval of the running variable that contains the two mass points, one on each side, that are immediately consecutive to the cutoff value. The key identifying assumption underlying this framework is that assignment of treatment is as good as random inside the interval. The

validity of this assumption can be easily tested using standard randomization tests which entails running series of regressions of observable student characteristics on the indicator for lower accountability rating.

Panel A of Appendix Table A8 presents these results for students in schools that received an Unsatisfactory or Below Average rating. The sample size in the interval that contains the two mass points is 2,695 and comes from 14 unique schools: 1,480 and 1,215 students attended schools that received an Unsatisfactory and Below Average rating, respectively. The discontinuity estimates are small in magnitude and none is statistically distinguishable from zero. Panel B reports the randomization test results for students in schools that received Below Average and Average rating (984 and 3,932 students in each mass point from a total of 24 schools, respectively).[17] Having shown convincing evidence on local randomization, we present the effects of accountability pressures from this alternative approach in Table 6. Reassuringly, the coefficient estimates are strikingly similar to those obtained using global polynomial fit (Table 5). The inference does not change in a meaningful way when we use the wild bootstrap t-procedure clustered at the school level to account for potential contamination in the inference procedure that may arise because of small number of schools (Cameron et al. 2008).

5.4 Mechanisms

The results from the previous sections indicate that accountability pressures, at the bottom end of the ratings distribution, decreased the arrest rates and improved economic self-sufficiency in adulthood. In this section, we consider potential explanations for these effects.

Could the results be attributable to student mobility? To the extent that lower accountability ratings led students to transfer out of their low-performing schools, students changing schools and moving to higher-quality schools may explain our results. To test this hypothesis, we created an indicator variable that takes the value one if the student switched schools in the academic years following accountability rating of their original school (between ninth and eleventh grades) and re-ran Equation (1) using this

[17] Note that there is some evidence for negative selection into schools rated Below Average in Panel B (Columns 2 and 3). Such selection is likely to bias the results from Panel B of Table 6 upward.

indicator as our outcome of interest. As shown in Column 1 of Table 7, we do not find any evidence of differential student mobility.

Existing studies of accountability also discuss the tendency of schools to manipulate the pool of test-takers by strategically exempting students from these tests (Cullen and Reback 2006; Figlio and Getzler 2006; Reback 2008; Neal and Schanzenbach 2010; Deming et al. 2016). This behavior of schools was largely motivated by the manner in which accountability systems were implemented in some states and districts where schools were assigned performance ratings based on the overall pass rates of eligible students in standardized tests. With the goal of boosting ratings, higher performing schools were more likely to classify low performing students as eligible for special education in order to exempt them from taking the high-stakes tests. Deming et al. (2016) show that, as a result of being placed in less-demanding academic tracks, these students ended up having worse educational and labor market outcomes. To investigate this potential mechanism, we created another indicator variable that takes the value one if a student received special education services in high school, although he or she had not received these services in eighth grade (Column 2). The discontinuity estimate from this exercise is not statistically different from zero, indicating that during the period we analyze in South Carolina there is no compelling evidence of strategic special education classification for schools at the bottom end of the ratings distribution. It is also important to note that the schools' scope for strategic Limited English Proficiency (LEP) classification is very limited in our context. This is because only around 1 percent of the analysis sample is flagged as LEP students.

Columns 3 and 4 of Table 7 present regression results where the dependent variable is grade progression. It bears noting that we do not have data on graduation status and as a result, we use information on grade progression to proxy for high school completion.[18] In these specifications, students are classified as being in the eleventh or twelfth grade if they had ever enrolled in these respective grades. The discontinuity estimates in these columns are statistically insignificant and the effects are almost equal to

[18] South Carolina's average on-time graduation in early 2000s was slightly below 60 percent (National Center for Education Statistics, 2005). SCDE provides information on high school completion beginning with the 2007-2008 academic year.

zero in magnitude. Taken together, these results suggest that the effects of accountability pressures on adult outcomes may operate through channels other than high school graduation.

Table 8 presents the discontinuity estimates related to the analysis of the relationship between accountability pressures and various features of educational production obtained from school report cards. These weighted regressions are run at the school-by-year level. There is no statistically significant impact of the receipt of a lower rating on teacher quality (measured by the fraction of teachers with an advanced degree) in Column 1, teacher turnover (measured by the fraction of teachers returning school from previous year) in Column 2, or per-pupil spending in Column 3, all of which are measured in the next academic year $(t+1)$ following the release of accountability ratings. Finally, we examined the relationship between the receipt of a lower accountability rating and schools' leadership change (Bacher-Hicks et al. 2019; Sorensen et al. 2022). More specifically, we created an indicator variable that takes the value one if a school's principal changed from one year to the next and used this measure as our outcome of interest. The discontinuity estimate, reported in the last column of Table 8, is negative (rather than positive) and it is not statistically different from zero indicating that a lower accountability rating did not lead to the replacement of the school's principal. Measuring these outcome variables in longer time horizon (e.g., $t+3$) do not change any of our findings.

Turning to non-financial processes of educational production, the results summarized in Table 9 show that the receipt of a lower rating increases schools' retention rates. At the same time, the percentage of tenth grade students meeting standards of the exit examination rises, and the same is true regarding the percentage of students eligible for merit-based (LIFE) scholarship (meeting both the GPA and SAT/ACT criteria) to a four-year institution, although this effect is not statistically significant at conventional levels (Columns 1-3). Taken together, the receipt of a lower rating appears to lead more students being retained, arguably because of increased standards, without having any net change in grade progression (Table 7) and the average student success, measured by exit exams and the eligibility for LIFE scholarship, rises.[19]

[19] We also examined the relationship between the receipt of a lower accountability rating and school's graduation rate. The estimated impact from this exercise is 1.409 (s.e.=3.604) and further supports the results reported in Table 7.

We also find an increase in student attendance rates–the point estimate is statistically significant at the 11% level and no meaningful change in suspension rates (Columns 4 and 5). Finally, to obtain an estimate of the impact on overall school outcomes and to reduce the chance of false positives (Kling et al. 2007), we created a school outcome index by averaging the z-scores of the variables from the first four columns. The point estimate for the school outcome index, reported in the last column, is positive and statistically significant at the 1% level. The receipt of a lower rating is associated with 0.50 of a standard deviation increase in school outcome index.

Receiving a lower rating and the associated accountability pressure appears to prompt schools to increase their academic standards and to implement procedures leading to enhanced academic success of their students. These changes are consistent with an explanation related to improvements in human capital accumulation.

6 Conclusion

School accountability systems are designed to evaluate the performance of public schools each year. With a portfolio of sanctions for low-performing schools and rewards for high performance, the goal of these accountability regimes is to incentivize schools to improve their students' academic outcomes. We analyze students and their schools which were exposed to South Carolina accountability regime, the implementation of which started in 2000. We link all students in the state to their records pertaining to interactions with the criminal justice system, and the welfare system up until they are in their early 30s.

We primarily focus on schools that are located at the low end of the ratings distribution, and therefore face more intense accountability pressures. We analyze students who are the first-time ninth graders in these schools in each year. Using a Regression Discontinuity framework, we find that lower-performing schools did not alter their average teacher quality, measured by the proportion of teachers with advanced degrees; nor did they change per pupil spending. Similarly, accountability pressures do not lead to a change in teacher turnover or leadership change at the school. There is no evidence for students

transferring out of these lower-rated schools; and we find that a school's receipt of a lower rating has no impact on its students' enrollment in subsequent grades.

We document that student academic performance increased in these lower-performing schools. Specifically, the proportion of students who passed the tenth grade exit exam increased. We also find a non-statistically significant rise in the proportion of students who qualify for merit-based college scholarships provided by the state. This increase in academic achievement is accompanied by a rise in academic standards, evidenced by an increase in the student retention rates. These findings indicate that low-performing schools responded to accountability pressures by increasing academic standards and improving student academic achievement.

These changes in the school environment had a positive impact on student's outcomes when they are adults. We find that students who attended lower-performing schools are less likely to engage in criminal activity by age 30, that they are less likely to be recipients of welfare benefits. These impacts are more pronounced for females. Further examination of the data also shows that the impact of participation in social welfare programs persists beyond early adulthood until the end of the data span, when individuals reach their 30s. We further complement our analysis using local randomization approach which changes the parameter of interest from the RD treatment effect at the cutoff to that in the immediate neighborhood around the cutoff where randomization holds. Finally, extending the analysis to schools that are at the top of the ratings distribution, we find no evidence of an effect. That is, whatever accountability pressures exist for the highly rated schools, they do not translate into a change in criminal involvement and economic self-sufficiency in adulthood. The linked administrative data we analyze do not contain information on individual earnings. However, participation in welfare programs such as food stamps and TANF is, by construction, determined by low income status, and research on economics of crime demonstrates the negative impact of criminal activity on wages, employment, and earnings. Thus, our results likely reflect an increase in earnings and decrease in joblessness during adulthood generated by a rise in human capital due to accountability pressures.

Improving low-performing schools is a perennial problem in education systems. Policymakers have implemented many strategies to turn around struggling schools. Our findings are intriguing in that they suggest the existence of policies and practices that low performing schools have implemented when they faced increased accountability pressures. A better understanding of specific educational responses is a useful area for future research as states transition to the Common Core standards.

References

Andrabi, T., J. Das, and A. I. Khwaja (2017). Report Cards: The Impact of Providing School and Child Test Scores on Educational Markets. *American Economic Review 107*(6), 1535–1563.

Autor, D., D. Figlio, K. Karbownik, J. Roth, and M. Wasserman (2019). Family Disadvantage and the Gender Gap in Behavioral and Educational Outcomes. *American Economic Journal: Applied Economics 11*(3), 338–381.

Bacher-Hicks, A., S. B. Billings, and D. J. Deming (2019). The School to Prison Pipeline: Long-Run Impacts of School Suspensions on Adult Crime. *NBER Working Paper 26257*.

Bald, A., E. Chyn, J. Hastings, and M. Machelett (2022). The Causal Impact of Removing Children from Abusive and Neglectful Homes. *Journal of Political Economy 130*(7), 1919–1962.

Bertrand, M. and J. Pan (2013). The Trouble with Boys: Social Influences and the Gender Gap in Disruptive Behavior. *American Economic Journal: Applied Economics 5*(1), 32–64.

Billings, S. B., D. J. Deming, and J. Rockoff (2014). School Segregation, Educational Attainment, and Crime: Evidence from the End of Busing in Charlotte-Mecklenburg. *Quarterly Journal of Economics 129*(1), 435–476.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). Bootstrap-Based Improvements for Inference with Clustered Errors. *Review of Economics and Statistics 90*(3), 414–427.

Carnoy, M. and S. Loeb (2002). Does External Accountability Affect Student Outcomes? A Cross-State Analysis. *Educational Evaluation and Policy Analysis 24*(4), 305–331.

Carson, A. E. (2014). Prisoners in 2013. *U.S. Department of Justice, Bureau of Justice Statistics*.

Cattaneo, M. D., N. Idrobo, and R. Titiunik (2018). A Practical Introduction to Regression Discontinuity Designs: Volume II. *Working Paper*.

Chiang, H. (2009). How Accountability Pressure on Failing Schools Affects Student Achievement. *Journal of Public Economics 93*(9), 1045–1057.

Cilliers, J., I. M. Mbiti, and A. Zeitlin (2021). Can Public Rankings Improve School Performance?: Evidence from a Nationwide Reform in Tanzania. *Journal of Human Resources 56*(3), 655–685.

Cullen, J. and R. Reback (2006). Tinkering Toward Accolades: School Gaming under a Performance Accountability System. In T. J. Gronberg and D. W. Jansen (Eds.), *Improving School Accountability*, Volume 14 of *Advances in Applied Microeconomics*, pp. 1–34. Emerald Group Publishing Limited.

Currie, J., J. Grogger, G. Burtless, and R. F. Schoeni (2001). Explaining Recent Declines in Food Stamp Program Participation. *Brookings-Wharton Papers on Urban Affairs*, 203–244.

Dee, T. S. and B. Jacob (2011). The Impact of No Child Left Behind on Student Achievement. *Journal of Policy Analysis and Management 30*(3), 418–446.

Deming, D. J., S. Cohodes, J. Jennings, and C. Jencks (2016). School Accountability, Postsecondary Attainment, and Earnings. *Review of Economics and Statistics 98*(5), 848–862.

Dizon-Ross, R. (2020). How Does School Accountability Affect Teachers?: Evidence from New York City. *Journal of Human Resources 55*(1), 76–118.

Figlio, D. N. (2006). Testing, Crime and Punishment. *Journal of Public Economics 90*(4), 837–851.

Figlio, D. N. and L. S. Getzler (2006). Accountability, Ability and Disability: Gaming the System? In T. J. Gronberg and D. W. Jansen (Eds.), *Improving School Accountability*, Volume 14 of *Advances in Applied Microeconomics*, pp. 35–49. Emerald Group Publishing Limited.

Figlio, D. N. and M. E. Lucas (2004). What's in a Grade? School Report Cards and the Housing Market. *American Economic Review 94*(3), 591–604.

Figlio, D. N. and C. E. Rouse (2006). Do Accountability and Voucher Threats Improve Low-Performing Schools? *Journal of Public Economics 90*(1), 239–255.

Figlio, D. N. and J. Winicki (2005). Food for Thought: The Effects of School Accountability Plans on School Nutrition. *Journal of Public Economics 89*(2), 381–394.

Frandsen, B. R. (2017). Party Bias in Union Representation Elections: Testing for Manipulation in the Regression Discontinuity Design when the Running Variable is Discrete. In *Regression Discontinuity Designs*, Volume 38 of *Advances in Econometrics*, pp. 281–315. Emerald Publishing Limited.

García, J. L., J. J. Heckman, and A. L. Ziff (2018). Gender Differences in the Benefits of An Influential Early Childhood Program. *European Economic Review 109*, 9–22.

Hanushek, E. A. and M. E. Raymond (2004). The Effect of School Accountability Systems on the Level and Distribution of Student Achievement. *Journal of the European Economic Association 2*(2-3), 406–415.

Kling, J. R., J. B. Liebman, and L. F. Katz (2007). Experimental Analysis of Neighborhood Effects. *Econometrica 75*(1), 83–119.

Ladd, H. F. (1999). The Dallas School Accountability and Incentive Program: An Evaluation of Its Impacts on Student Outcomes. *Economics of Education Review 18*(1), 1–16.

Neal, D. and D. W. Schanzenbach (2010). Left Behind by Design: Proficiency Counts and Test-Based Accountability. *Review of Economics and Statistics 92*(2), 263–283.

Nunes, L. C., A. B. Reis, and C. Seabra (2015). The Publication of School Rankings: A Step toward Increased Accountability? *Economics of Education Review 49*, 15–23.

Reback, R. (2008). Teaching to the Rating: School Accountability and the Distribution of Student Achievement. *Journal of Public Economics 92*(5), 1394–1415.

Reback, R., J. Rockoff, and H. L. Schwartz (2014). Under Pressure: Job Security, Resource Allocation, and Productivity in Schools under No Child Left Behind. *American Economic Journal: Economic Policy 6*(3), 207–241.

Rockoff, J. and L. J. Turner (2010). Short-Run Impacts of Accountability on School Quality. *American Economic Journal: Economic Policy 2*(4), 119–147.

Rouse, C. E., J. Hannaway, D. Goldhaber, and D. Figlio (2013). Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure. *American Economic Journal: Economic Policy 5*(2), 251–281.

Sorensen, L. C., S. D. Bushway, and E. J. Gifford (2022). Getting Tough? The Effects of Discretionary Principal Discipline on Student Outcomes. *Education Finance and Policy 17*(2), 255–284.

South Carolina Education Oversight Committee (2003). Accountability Manual (2000-2003).

Table I: Summary Statistics by Accountability Ratings

		School Ratings				
	All	**Unsatisfactory**	**Below Average**	**Average**	**Good**	**Excellent**
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Student Characteristics						
Black	0.414	0.792	0.671	0.554	0.363	0.258
White	0.554	0.191	0.297	0.417	0.608	0.698
Female	0.481	0.471	0.465	0.479	0.480	0.485
Free/Reduced Lunch	0.406	0.677	0.595	0.503	0.404	0.254
Proficient in Math-8th Grade	0.337	0.302	0.274	0.247	0.351	0.379
Proficient in ELA-8th Grade	0.374	0.325	0.290	0.282	0.381	0.432
Panel B: Adult Outcomes						
Adult Arrest	0.199	0.239	0.238	0.216	0.204	0.167
Adult Incarceration	0.052	0.081	0.077	0.059	0.052	0.037
Participation in Food Stamps as an Adult	0.513	0.738	0.667	0.596	0.526	0.370
Participation in TANF as an Adult	0.122	0.203	0.173	0.146	0.121	0.081
Welfare Participation	0.515	0.740	0.672	0.597	0.527	0.372
Sample Size	161,281	13,365	13,932	19,074	62,445	52,465
Panel C: School Characteristics						
Percent Teachers with an Advanced Degree	49.26	42.94	42.92	45.08	49.47	55.99
Percent Teachers Returning School from Previous Year	84.25	79.49	82.11	82.08	85.63	86.19
Professional Development Days (per year) for Teachers	9.29	9.07	9.55	8.83	9.32	9.52
Per Pupil Spending	6064.82	6693.23	6109.42	6423.61	5876.13	5864.05
Number of School-Year Observations	549	70	52	74	202	151

NOTES: The tabulations reflect our research sample which comprises three cohorts of first-time ninth graders in public high schools between the 2000-2001 and 2002-2003 academic years. A student performing at or above the Proficient level on the state's eighth grade subject-specific assessments is labeled as proficient. The full set of sample statistics is available from authors upon request.

Table 2: Regression Discontinuity Validation Tests

	Female	Free Lunch	White	Age First Found in Public School	Proficient in 8th Grade Math	Proficient in 8th Grade ELA
	Coefficients (Standard Errors)					
	(1)	(2)	(4)	(5)	(6)	(7)
Panel A: Full Sample						
Receipt of Lower Rating	-0.003	-0.010	0.048	-0.027	0.019	0.009
	(0.008)	(0.042)	(0.053)	(0.069)	(0.014)	(0.013)
Sample Size	46,371	46,371	46,371	46,371	33,871	33,467
Panel B: Females						
Receipt of Lower Rating	…..	-0.016	0.048	-0.028	0.014	0.022
		(0.042)	(0.053)	(0.068)	(0.015)	(0.015)
Sample Size		21,935	21,935	21,935	16,903	16,385
Panel C: Males						
Receipt of Lower Rating	…..	-0.005	0.048	-0.029	0.023	-0.003
		(0.044)	(0.054)	(0.074)	(0.015)	(0.015)
Sample Size		24,436	24,436	24,436	16,968	17,082

NOTES: Standard errors are clustered at the school level. All specifications control for a quartic in school's accountability score, cohort fixed effects and interactions of cohort fixed effects with the quartic accountability score. The outcome variables in Columns (6) and (7) take the value one if the student performed at or above the Proficient level on the state's eighth grade subject-specific assessments. Receipt of a lower rating is an indicator denoting a lower accountability rating from the bottom thresholds together (Average/Below Average and Below Average/Unsatisfactory).

Table 3: Regression Discontinuity Estimates of the Effect of Accountability Ratings on Adult Crime

	Adult Arrest					Adult Incarc.
				Females	Males	
	Coefficients (Standard Errors)					
	(1)	(2)	(3)	(4)	(5)	(6)
Receipt of Lower Rating	-0.018**	-0.017**	-0.018**	-0.020***	-0.016	-0.003
	(0.009)	(0.008)	(0.009)	(0.007)	(0.013)	(0.005)
Control Mean	0.223			0.150	0.291	0.069
Sample Mean	46,371	46,371	46,371	21,935	24,436	46,371
Controls:						
Cohort Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Student Characteristics	No	Yes	Yes	Yes	Yes	Yes
School Characteristics	No	No	Yes	Yes	Yes	Yes

NOTES: Standard errors are clustered at the school level. All specifications control for a quartic in school's accountability score, cohort fixed effects and interactions of cohort fixed effects with the quartic accountability score. Student level controls include indicators for gender, race, free/reduced lunch status and age student was first found in public school. School characteristics include the percent of ninth-graders who are female, black, free/reduced lunch eligible and average age first found in public school. Adult crime takes the value one if individual was ever arrested as an adult in Columns (1)-(5) and it takes the value one if individual was ever incarcerated as an adult in Column (6). Receipt of a lower rating is an indicator denoting a lower accountability rating from the bottom thresholds together (Average/Below Average and Below Average/Unsatisfactory).

*** significant at 1%, ** significant at 5%.

Table 4: Regression Discontinuity Estimates of the Effect of Accountability Ratings on Economic Self-Sufficiency

	Full Sample	Females	Males
	Coefficients (Standard Errors)		
	(1)	(2)	(3)
Panel A: Welfare Participation			
Receipt of Lower Rating	-0.028	-0.042**	-0.017
	(0.020)	(0.020)	(0.021)
Control Mean	0.622	0.699	0.552
Panel B: Food Stamps			
Receipt of Lower Rating	-0.029	-0.043**	-0.017
	(0.020)	(0.021)	(0.021)
Control Mean	0.621	0.698	0.551
Panel A: TANF			
Receipt of Lower Rating	-0.008	-0.018	0.001
	(0.010)	(0.018)	(0.005)
Control Mean	0.155	0.265	0.055
Sample Size	46,371	21,935	24,436

NOTES: Standard errors are clustered at the school level. All specifications control for a quartic in school's accountability score, cohort fixed effects and interactions of cohort fixed effects with the quartic accountability score. Covariates include indicators for gender, race, free/reduced lunch status, age student was first found in public school, the percent of ninth-graders who are female, black, free/reduced lunch eligible and average age first found in public school. Welfare participation takes the value one if individual was ever enrolled in social programs (food stamps /SNAP and TANF) as an adult. Receipt of a lower rating is an indicator denoting a lower accountability rating from the bottom thresholds together (Average/Below Average and Below Average /Unsatisfactory).

** significant at 5%.

Table 5: Regression Discontinuity Estimates of the Effect of Accountability Ratings on Long-Run Outcomes-Each Separate Cutoffs

	Adult Crime	Welfare Part. Full Sample	Welfare Part. Females
		Coefficients (Standard Errors)	
	(1)	(2)	(3)
Accountability Rating			
Unsatisfactory	-0.095**	-0.015	-0.077
	(0.039)	(0.082)	(0.084)
Below Average	-0.029	-0.015	-0.041
	(0.024)	(0.056)	(0.058)
p-value-Test of Equal Coefficients ($\beta_U = \beta_{BA}$)	0.00	0.98	0.29
Sample Size	46,371	46,371	21,935

NOTES: Standard errors are clustered at the school level. All specifications control for a quartic in school's accountability score, cohort fixed effects and interactions of cohort fixed effects with the quartic accountability score. Covariates include indicators for gender, race, free/reduced lunch status, age student was first found in public school, the percent of ninth-graders who are female, black, free/reduced lunch eligible and average age first found in public school.
** significant at 5%.

Table 6: Estimates of the Effect of Accountability Ratings on Long-Run Outcomes-Local Randomization Approach

	Adult Crime	Welfare Part. Full Sample	Welfare Part. Females
		Coefficients (Standard Errors)	
	(1)	(2)	(4)
Panel A: Unsatisfactory vs. Below Average (N=2,695)	-0.078***	-0.016	-0.030
Receipt of Lower Rating	(0.019)	(0.035)	(0.041)
Sample Size	2,695	2,695	1,276
Panel B: Below Average vs. Average (N=4,916)			
Receipt of Lower Rating	-0.017	0.014	-0.021
	(0.020)	(0.049)	(0.057)
Sample Size	4,916	4,916	2,362

NOTES: The analysis sample is restricted to the interval of the running variable that contains only the two mass points, one on each side, that are immediately consecutive to the cutoff value. Standard errors are clustered at the school level. Covariates include cohort fixed effects, indicators for gender, race, free/reduced lunch status, age student was first found in public school, the percent of ninth-graders who are female, black, free/reduced lunch eligible and average age first found in public school level. Receipt of a lower rating in Panel A (Panel B) is an indicator denoting lower accountability rating for students in schools that received Unsatisfactory or Below Average (Below Average or Average) rating.

Table 7: Mechanisms-Regression Discontinuity Estimates of the Effect of Accountability Ratings on Mobility, Special Education and School Enrollment

	Changed School	Classified as Special Educ.	Enrolled in 11th Grade	Enrolled in 12th Grade
	Coefficients (Standard Errors)			
	(1)	(2)	(3)	(4)
Receipt of Lower Rating	-0.008	-0.011	0.005	0.009
	(0.013)	(0.011)	(0.021)	(0.023)
Control Mean	0.062	0.034	0.603	0.581
Sample Size	46,371	37,563	46,371	46,371

NOTES: Standard errors are clustered at the school level. All specifications control for a quartic in school's accountability score, cohort fixed effects and interactions of cohort fixed effects with the quartic accountability score. Covariates include indicators for gender, race, free/reduced lunch status, age student was first found in public school, the percent of ninth-graders who are female, black, free/reduced lunch eligible and average age first found in public school. The dependent variable in Column (1) takes the value one if student ever changed school between ninth and eleventh grades, while, in Column (2), it takes the value one if student was classified as special education in high school, but had not received special education services in middle school. Receipt of a lower rating is an indicator denoting a lower accountability rating from the bottom thresholds together (Average/Below Average and Below Average/Unsatisfactory).

Table 8: Mechanism-Regression Discontinuity Estimates of the Effect of Accountability Ratings on School Characteristics

	% Teachers with an Advanced Degree	% Teachers Returning School from Previous Year	Per Pupil Spending	Leadership Change
	Coefficients (Standard Errors)			
	(1)	(2)	(3)	(4)
Receipt of Lower Rating	2.571	0.706	-67.18	-0.032
	(2.246)	(1.789)	(300.32)	(0.124)
Control Mean	45.43	82.63	6,665.30	0.260
Sample Size	179	177	178	187

NOTES: Standard errors are clustered at the school level. All specifications control for a quartic in school's accountability score, cohort fixed effects score, cohort fixed effects and interactions of cohort fixed effects with the quartic accountability score. All outcomes are measured at the school-year level using aggregate information from *(t+1)*. Regressions are weighted by the total number of teachers in Columns 1 and 2 and by the total school enrollment in Column 3. Covariates include the percent of students who are female, black, free/reduced lunch eligible and average age first found in who are female, black, free/reduced lunch eligible and average age first found in public school. The dependent variable in Column (4) takes the value one if school's principal changed from *(t)* to *(t+1)*. Receipt of a lower rating is an indicator denoting a lower accountability rating from the bottom thresholds together (Average/Below Average and Below Average/Unsatisfactory).
***significant at 1%; ** significant at 5%.

Table 9: Mechanism-Regression Discontinuity Estimates of the Effect of Accountability Ratings on School Outcomes

	% Students Retained	% 10th Grade Students Passing the Exit Exams	% Students Eligible for LIFE Scholarship	Student Attendance Rate	% Students Susp./Expelled	School Outcome Index-I
	Coefficients (Standard Errors)					
	(1)	(2)	(3)	(4)	(5)	(6)
Receipt of Lower Rating	4.090**	4.553**	1.234	1.029	-0.219	0.500***
	(1.594)	(2.066)	(1.803)	(0.633)	(1.975)	(0.162)
Control Mean	9.30	62.08	8.27	95.39	4.47	0.101
Sample Size	179	183	184	187	123	175

NOTES: Standard errors are clustered at the school level. All specifications control for a quartic in school's accountability score, cohort fixed effects and interactions of cohort fixed effects with the quartic accountability score. All outcomes are measured at the school-year level. Regressions are weighted by the total school enrollment. Covariates include the percent of students who are female, black, free/reduced lunch eligible and average age first found in public school. Eligibility for LIFE scholarships to a four-year institution is based on meeting both the grade point average and SAT/ACT criteria established by the state. School Index-I includes the outcomes from Columns (1)-(4). The indix is constructed by averaging *z-scores* of each component. Receipt of a lower rating is an indicator denoting a lower accountability rating from the bottom thresholds together (Average/Below Average and Below Average/Unsatisfactory).

***significant at 1%; ** significant at 5%.

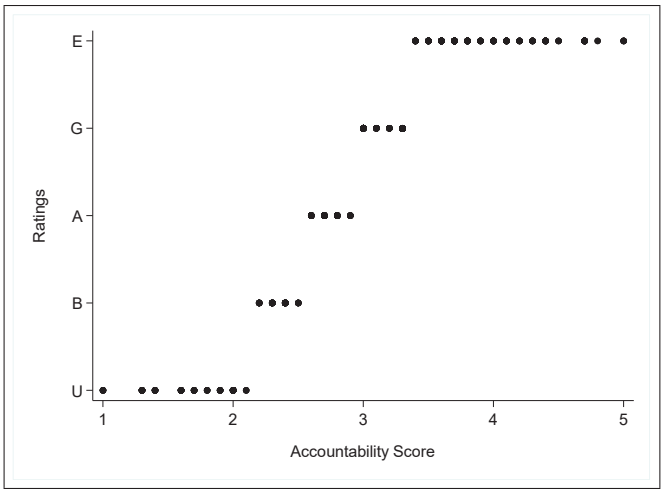
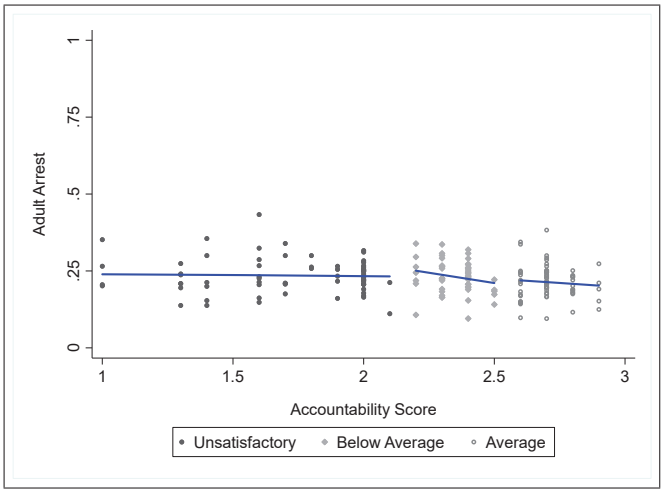


Figure 1: Distribution of Schools by Accountability Ratings

NOTES: The figure displays the distribution of schools by accountability ratings between the 2000-2001 and 2002-2003 academic years. Schools were assigned one of five performance ratings: (i) Unsatisfactory (U), (ii) Below Average (B), (iii) Average (A), (iv) Good (G), and (v) Excellent (E).

Panel A: Adult Arrest



Panel B: Welfare Participation

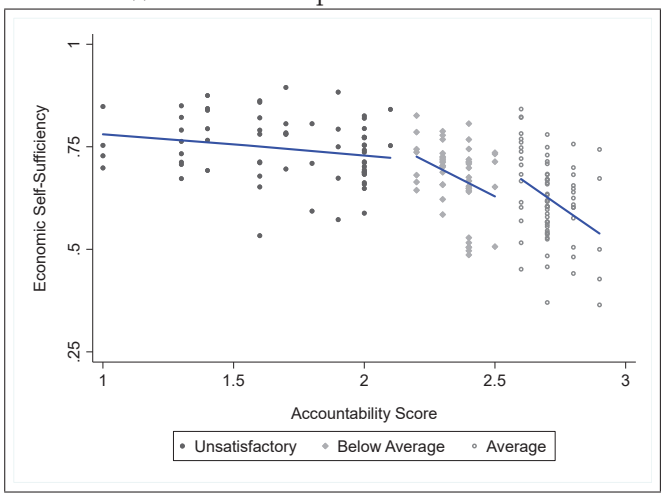
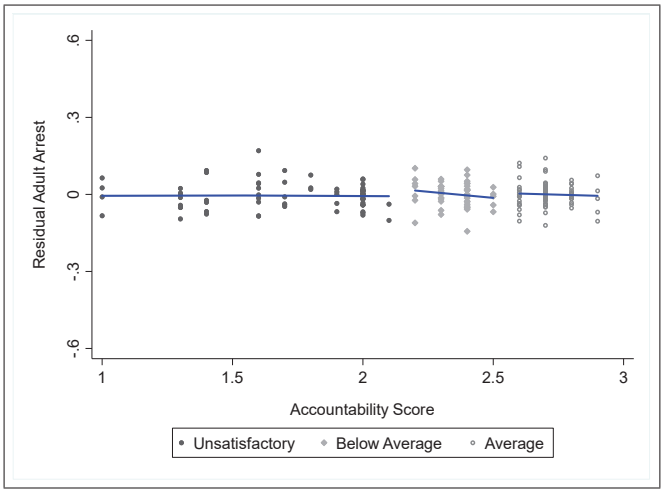


Figure 2: Raw Long-Run Outcomes and Accountability Ratings

NOTES: The solid lines are estimates from locally weighted polynomial regressions.

Panel A: Adult Arrest



Panel B: Welfare Participation

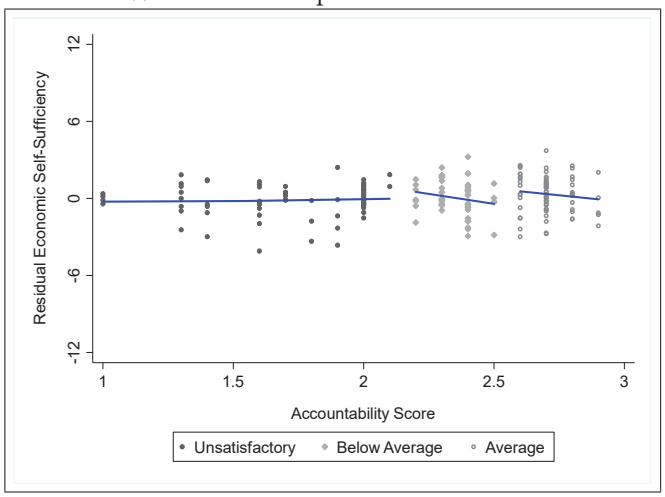


Figure 3: Residualized Long-Run Outcomes and Accountability Ratings

NOTES: Residuals in Panel A (Panel B) are obtained from a regression of school's average adult arrest (welfare participation) rate on a quartic in accountability score, cohort fixed effects, interactions of cohort fixed effects with the quartic accountability score and school level controls (percent of ninth graders who were female, black, free/reduced lunch eligible and average age first found in public school). Regressions are weighted by the number of ninth graders at the school. The solid lines are estimates from locally weighted polynomial regressions.

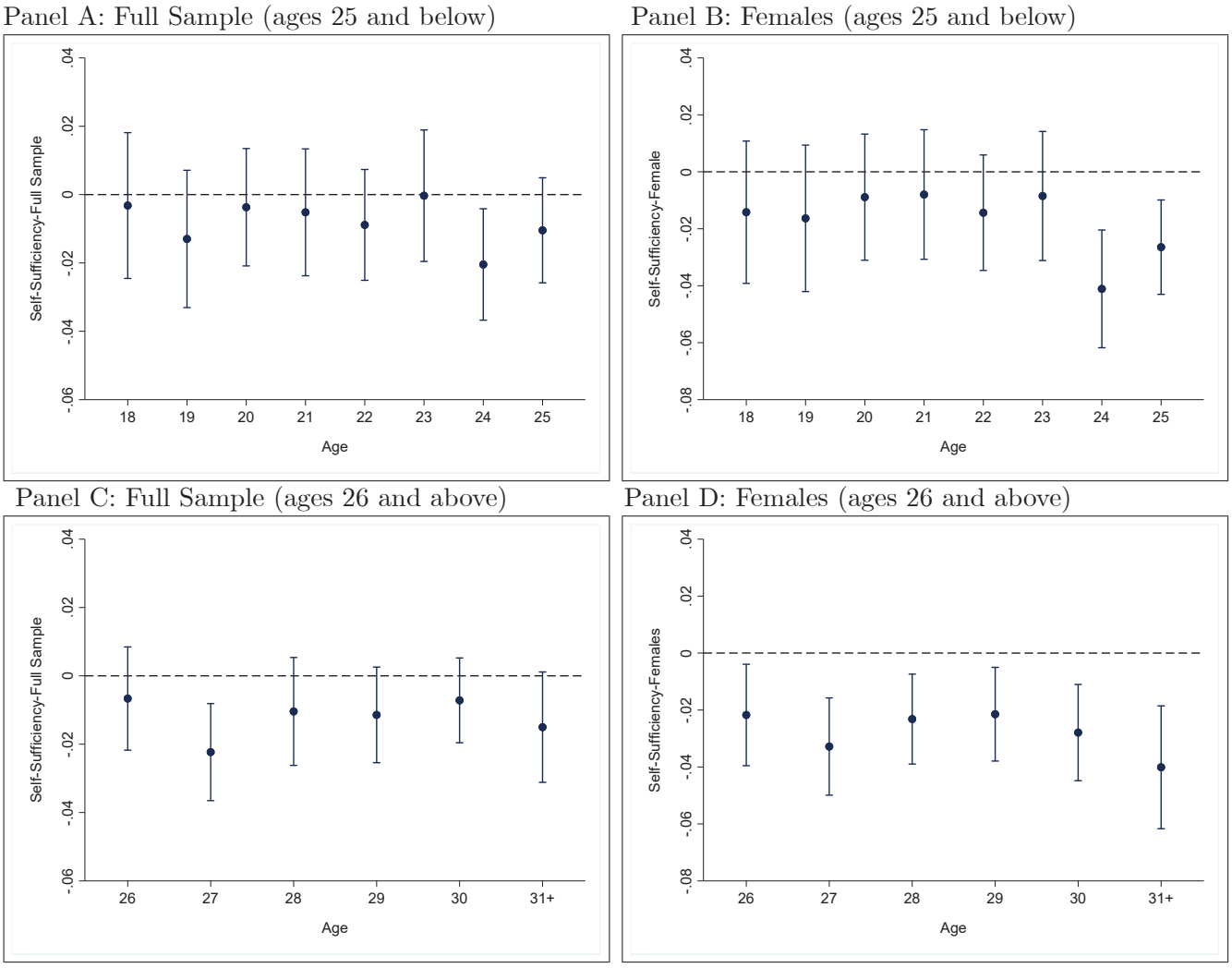
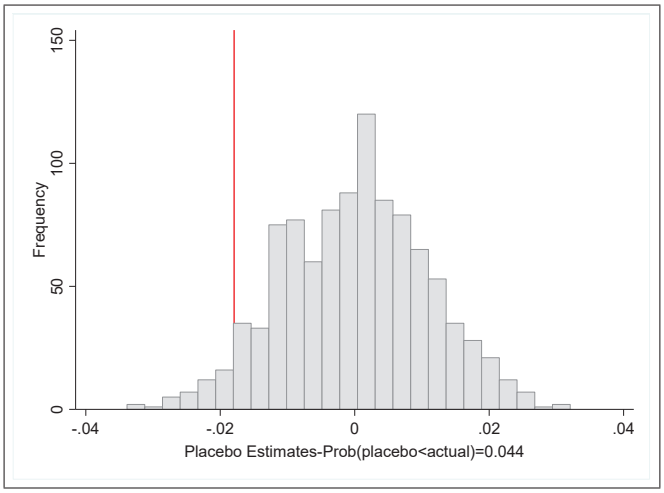


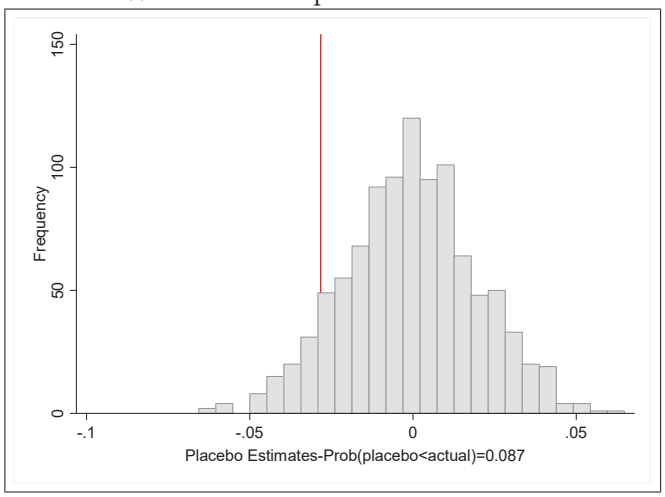
Figure 4: The Effect of Receiving a Lower Accountability Rating on Welfare Participation-by Age

NOTES: Each point in each panel comes from a separate regression, using samples that increase in age moving rightward along the x-axis. The dependent variable takes the value one if individual enrolled in social programs (food stamps/SNAP or TANF) by the given age. Each dot represents the regression discontinuity coefficient, obtained by equation (1). The height of the bars extending from each point represents the bounds of the 90% confidence interval.

Panel A: Adult Crime



Panel B: Welfare Participation



Panel C: Welfare Participation-Females

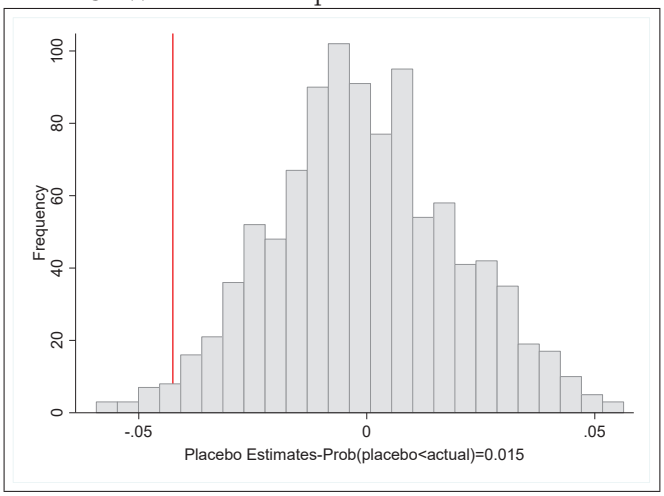


Figure 5: Placebo Coefficients of the Effect of Accountability Ratings

NOTES: The figure displays the distribution of placebo coefficients of the effect of accountability ratings, where the accountability scores for a given year are randomly assigned to different schools. The vertical line represents the actual point estimate reported in Tables 3 and 4.

Table A1: Regression Discontinuity Validation Tests-School Characteristics

	%Female	%Free Lunch	%White	Average Age First Found in Public School	% Teachers with an Advanced Degree	% Teachers Returning School from Previous Year	Per Pupil Spending
				Coefficients (Standard Errors)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Receipt of Lower Rating	-0.317	-1.869	4.888	0.001	1.634	1.116	-114.68
	(0.718)	(4.226)	(5.244)	(0.027)	(2.050)	(1.594)	(385.21)
Sample Size	196	196	196	196	187	182	184

NOTES: Standard errors are clustered at the school level. All specifications control for a quartic in school's accountability score, cohort fixed effects and interactions of cohort fixed effects with the quartic accountability score. Regressions in Columns (1)-(4) and (7) are weighted by the total school enrollment, while those in Columns (5) and (6) are weighted by the total number of teachers. Receipt of a lower rating is an indicator denoting a lower accountability rating from the bottom thresholds together (Average/Below Average and Below Average/Unsatisfactory).

Table A2: Regression Discontinuity Estimates of the Effect of Accountability Ratings on Long-Run Outcomes-by Age

	Age<=25	Age>25
	Coefficients (Standard Errors)	
	(1)	(2)
Panel A: Adult Crime		
Receipt of Lower Rating	-0.015	-0.010
	(0.010)	(0.007)
Control Mean	0.195	0.117
Sample Size	46,371	46,371
Panel B: Welfare Participation		
Receipt of Lower Rating	-0.023	-0.025
	(0.019)	(0.017)
Control Mean	0.563	0.430
Sample Size	46,371	46,371
Panel C: Welfare Participation-Females		
Receipt of Lower Rating	-0.035*	-0.058***
	(0.019)	(0.018)
Control Mean	0.646	0.535
Sample Size	21,935	21,935

NOTES: Standard errors are clustered at the school level. All specifications control for a quartic in school's accountability score, cohort fixed effects and interactions of cohort fixed effects with the quartic accountability score. Covariates include indicators for gender, race, free/reduced lunch status, age student was first found in public school, the percent of ninth-graders who are female, black, free/reduced lunch eligible and average age first found in public school. The dependent variable in Column (1) of Panel A takes the value of one if individual was 25 years old or below at the time of offense (welfare participation in Panels B and C). The dependent variable in Column (2) is defined similarly. Receipt of a lower rating is an indicator denoting a lower accountability rating from the bottom thresholds together (Average/Below Average and Below Average/ Unsatisfactory).
** significant at 5%.

Table A3: Robustness Checks-Regression Discontinuity Estimates of the Effect of Accountability Ratings on Long-Run Outcomes: Alternative Specifications and Bandwidths

	Cubic in Accountability Score	Cubic in Account. Score Distance to Cutoff=[-0.6,.06]	Quadratic Account. Score Distance to Cutoff=[-0.5,0.5]	Local Linear Distance to Cutoff=[-0.5,0. 5]
	Coefficients (Standard Errors)			
	(1)	(2)	(3)	(4)
Panel A: Adult Crime				
Receipt of Lower Rating	-0.017*	-0.019**	-0.017*	-0.026**
	(0.009)	(0.009)	(0.009)	(0.010)
Sample Size	46,371	43,260	41,809	41,809
Panel B: Welfare Participation				
Receipt of Lower Rating	-0.027	-0.028	-0.026	-0.010
	(0.020)	(0.020)	(0.020)	(0.010)
Sample Size	46,371	43,260	41,809	41,809
Panel C: Welfare Participation-Females				
Receipt of Lower Rating	-0.043**	-0.042**	-0.042**	-0.034**
	(0.021)	(0.021)	(0.021)	(0.013)
Sample Size	21,935	20,528	19,874	19,874

Table A3 cont. Robustness Checks-Regression Discontinuity Estimates of the Effect of Accountability Ratings on Long-Run Outcomes: Alternative Specifications and Bandwidths

	Add 8th Grade Performance Indicators	Add More Recent Cohorts	Donut RD	School Level (Weighted)	Placebo Test-Future Ratings Assignment
	Coefficients (Standard Errors)				
	(5)	(6)	(7)	(8)	(9)
Panel A: Adult Crime					
Receipt of Lower Rating	-0.017**	-0.012**	-0.019*	-0.017*	0.009
	(0.008)	(0.006)	(0.010)	(0.008)	(0.009)
Sample Size	46,371	99,304	42,439	196	33,311
Panel B: Welfare Participation					
Receipt of Lower Rating	-0.024	-0.019	-0.037*	-0.025	0.015
	(0.019)	(0.012)	(0.021)	(0.019)	(0.022)
Sample Size	46,371	99,304	42,439	196	33,311
Panel C: Welfare Participation-Females					
Receipt of Lower Rating	-0.036**	-0.026**	-0.046**	-0.041**	0.023
	(0.018)	(0.011)	(0.022)	(0.019)	(0.022)
Sample Size	21,935	47,682	20,045	196	33,311

NOTES: Standard errors are clustered at the school level. Column (1) controls for a cubic in school's accountability score. Column (2) controls for a cubic in school's accountability score and limits the sample to schools whose scores were within 0.6 points of F or D discontinuities. Column (3) controls for a quadratic in school's accountability score and limits the sample to schools whose scores were within 0.5 points of F or D discontinuities, while Column (4) reports the results from a local linear specification by limiting the sample to within 0.5 points. Column (5) adds eighth grade subject-specific proficiency indicators to the baseline specifications. A separate indicator for missing value in eighth grade proficiency measures is also included. Column (6) extends the data to include more recent ninth grade cohorts (2000-2001 to 2005-2006 academic years). Column (7) repeats the results using donut-RD models where we remove schools with 2.2 and 2.6 points. Column (8) collapses the data at the school level, weighted by the number of ninth graders at the school. Finally, Column (9) assigns schools accountability ratings from (t+4). Covariates include indicators for gender, race, free/reduced lunch status, age student was first found in public school, the percent of ninth graders who are female, black, free/reduced lunch eligible and average age first found in public school. Receipt of a lower rating is an indicator denoting a lower accountability rating from the bottom thresholds together (Average/Below Average and Below Average/Unsatisfactory).
** significant at 5%; * significant at 10%.

Table A4: Regression Discontinuity Validation Tests Including More Recent Cohorts (2000-2001 to 2005-2006 academic years)

	Female	Free Lunch	White	Age First Found in Public School	Proficient in 8th Grade Math	Proficient in 8th Grade ELA
	Coefficients (Standard Errors)					
	(1)	(2)	(4)	(5)	(6)	(7)
Receipt of Lower Rating	-0.004	-0.004	0.028	-0.024	0.005	0.001
	(0.005)	(0.026)	(0.039)	(0.049)	(0.009)	(0.008)
Sample Size	99,304	99,304	99,304	99,304	80,496	80,091

NOTES: Standard errors are clustered at the school level. All specifications control for a quartic in school's accountability score, cohort fixed effects and interactions of cohort fixed effects with the quartic accountability score. The outcome variables in Columns (6) and (7) take the value one if the student performed at or above the Proficient level on the state's eighth grade subject-specific assessments. Receipt of a lower rating is an indicator denoting a lower accountability rating from the bottom thresholds together (Average/Below Average and Below Average/Unsatisfactory).

Table A5: Regression Discontinuity Estimates of the Effect of Accountability Ratings on Long-Run Outcomes-by Student's Proficiency Level in Eighth Grade Standardized Tests

	Proficient in 8th Grade Math or ELA Subject Tests	Below Proficient in 8th Grade Math and ELA Subject Tests
	Coefficients (Standard Errors)	
	(1)	(2)
Panel A: Adult Crime		
Receipt of Lower Rating	-0.018	-0.006
	(0.013)	(0.009)
Control Mean	0.169	0.238
Sample Size	11,329	21,680
Panel B: Welfare Participation		
Receipt of Lower Rating	-0.011	-0.016
	(0.018)	(0.019)
Control Mean	0.511	0.692
Sample Size	11,329	21,680
Panel C: Welfare Participation-Females		
Receipt of Lower Rating	-0.008	-0.023
	(0.022)	(0.019)
Control Mean	0.583	0.791
Sample Size	6,155	10,076

NOTES: Standard errors are clustered at the school level. All specifications control for a quartic in school's accountability score, cohort fixed effects and interactions of cohort fixed effects with the quartic accountability score. Covariates include indicators for gender, race, free/reduced lunch status, age student was first found in public school, the percent of ninth-graders who are female, black, free/reduced lunch eligible and average age first found in public school.

Table A6: Regression Discontinuity Validation Tests-Top End of the Ratings Distribution

	Female	Free Lunch	White	Age First Found in Public School	Proficient in 8th Grade Math	Proficient in 8th Grade ELA
	Coefficients (Standard Errors)					
	(1)	(2)	(4)	(5)	(6)	(7)
Receipt of Lower Rating	-0.003	-0.050**	0.059**	0.008	0.027***	0.012
	(0.004)	(0.020)	(0.023)	(0.021)	(0.010)	(0.009)
Sample Size	133,359	133,359	133,359	133,359	99,786	98,375

NOTES: Standard errors are clustered at the school level. All specifications control for a quartic in school's accountability score, cohort fixed effects and interactions of cohort fixed effects with the quartic accountability score. The outcome variables in Columns (6) and (7) take the value one if the student performed at or above the Proficient level on the state's eighth grade subject-specific assessments. Receipt of a lower rating is an indicator denoting a lower accountability rating from the top thresholds together (Excellent/Good and Good/Average).

Table A7: Regression Discontinuity Estimates of the Effect of Accountability Ratings on Long-Run Outcomes-Top End of the Ratings Distribution

	Adult Crime	Welfare Part. Full Sample	Welfare Part. Females
		Coefficients (Standard Errors)	
	(1)	(2)	(3)
Receipt of Lower Rating	-0.004	-0.001	0.007
	(0.004)	(0.009)	(0.009)
Control Mean	0.184	0.449	0.507
Sample Mean	133,359	133,359	66,334

NOTES: Standard errors are clustered at the school level. All specifications control for a quartic in school's accountability score, cohort fixed effects and interactions of cohort fixed effects with the quartic accountability score. Covariates include indicators for gender, race, free/reduced lunch status, age student was first found in public school, the percent of ninth-graders who are female, black, free/reduced lunch eligible and average age first found in public school. Receipt of a lower rating is an indicator denoting a lower accountability rating from the top thresholds together (Excellent/Good and Good/Average).

Table A8: Local Randomization Tests

	Female	Free Lunch	White	Age First Found in Public School	Proficient in 8th Grade Math	Proficient in 8th Grade ELA
	Coefficients (Standard Errors)					
	(1)	(2)	(4)	(5)	(6)	(7)
Panel A: Unsatisfactory vs. Below Average (N=2,695)						
Receipt of Lower Rating	-0.014	0.022	-0.028	0.084	0.026	-0.013
	(0.019)	(0.079)	(0.080)	(0.067)	(0.024)	(0.026)
Panel B: Below Average vs. Average (N=4,916)						
Receipt of Lower Rating	-0.007	0.106	-0.089	0.018	0.011	-0.034
	(0.017)	(0.067)	(0.079)	(0.067)	(0.028)	(0.030)

NOTES: The analysis sample is restricted to the interval of the running variable that contains only the two mass points, one on each side, that are immediately consecutive to the cutoff value. Standard errors are clustered at the school level. All specifications control for cohort fixed effects. The outcome variables in Columns (6) and (7) take the value one if the student performed at or above the Proficient level on the state's eighth grade subject-specific assessments. Receipt of a lower rating in Panel A (Panel B) is an indicator denoting a lower accountability rating for students in schools that received an Unsatisfactory or Below Average (Below Average or Average) rating.

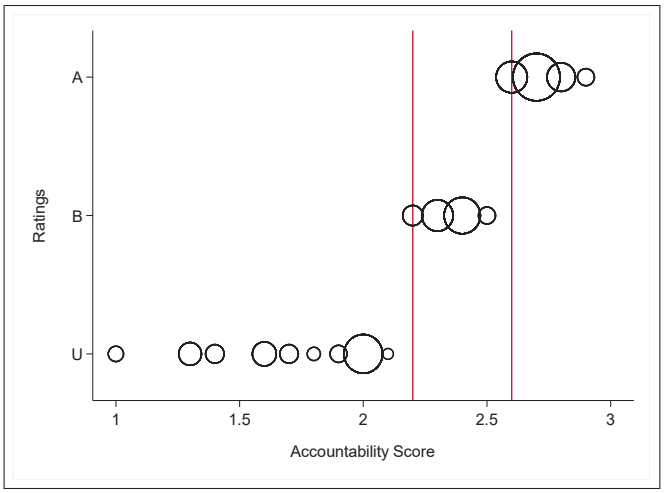


Figure A1: Distribution of Schools at the Bottom End of the Ratings Distribution

NOTES: The number of schools at each accountability score is proportional to the size of the bubble. The vertical lines denote the actual cutoffs for receiving a Below Average and Average ratings, respectively.

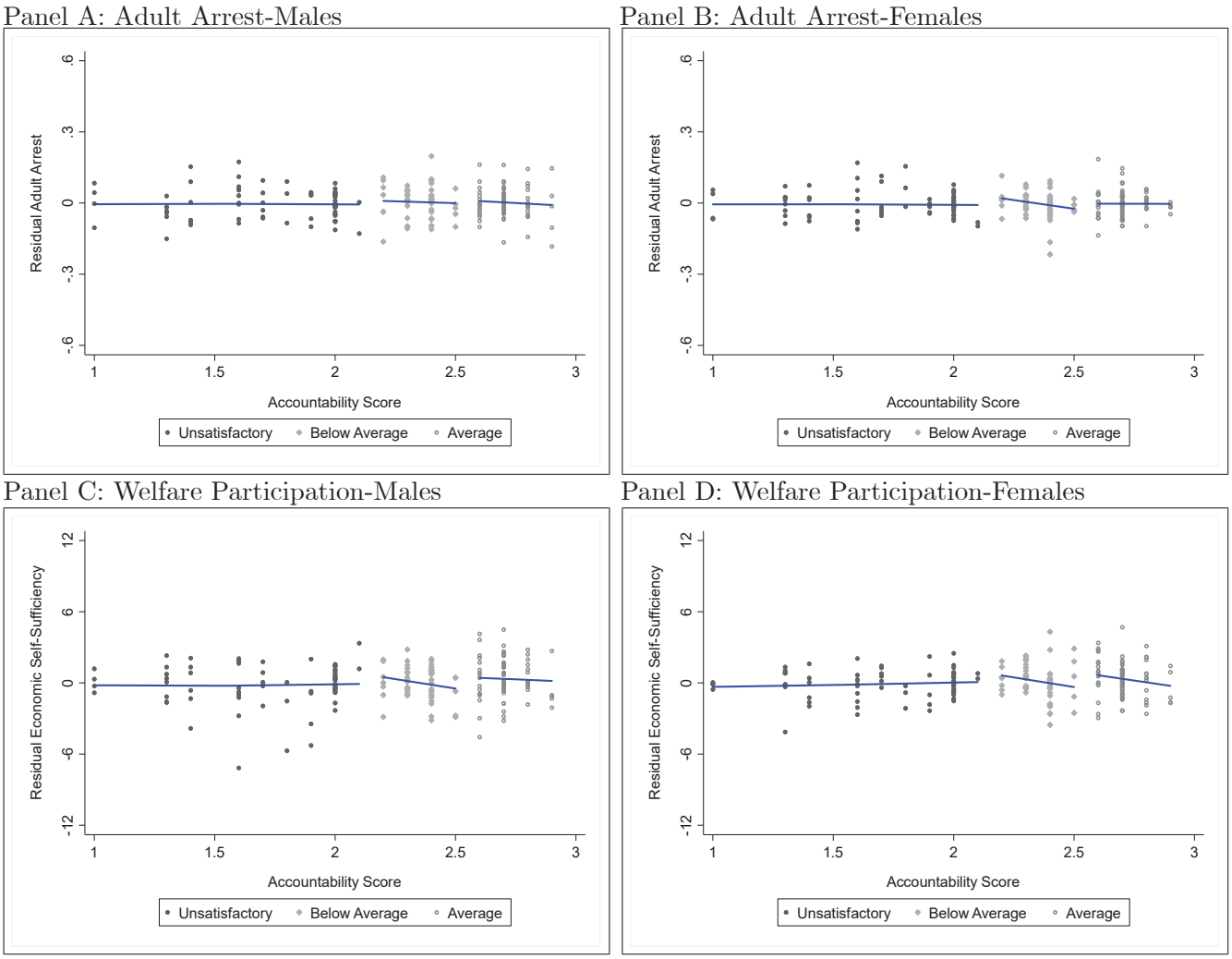


Figure A2: Residualized Long-Run Outcomes and Accountability Ratings-by Gender

NOTES: Residuals in Panels A and B (Panels C and D) are obtained from regressions of school's average gender-specific adult arrest (welfare participation) rate on a quartic in accountability score, cohort fixed effects, interactions of cohort fixed effects with the quartic accountability score and school level controls (percent of ninth graders who were female, black, free/reduced lunch eligible and average age first found in public school). Regressions are weighted by the number of ninth graders at the school. The solid lines are estimates from locally weighted polynomial regressions.