EFFECTS OF REWARD-BASED SCHOOL ACCOUNTABILITY SYSTEMS

Simon Calmar Andersen
Department of Political Science
TrygFonden's Centre for Child Research
Aarhus University
Bartholin salle 7
8000 Aarhus C
Denmark
sca@ps.au.dk

Ulrik Hvidman
Department of Political Science
TrygFonden's Centre for Child Research
Aarhus University
Bartholin salle 7
8000 Aarhus C
Denmark
uhvidman@ps.au.dk

Sarah Yde Junge
Department of Political Science
TrygFonden's Centre for Child Research
Aarhus University
Bartholin salle 7
8000 Aarhus C
Denmark
syj@ps.au.dk

Beatrice Schindler Rangvid
The Danish Centre for Social Science Reserach (VIVE)
Herluf Trolles Gade 11
1052 Copenhagen K
Denmark
bsr@vive.dk

EFFECTS OF REWARD-BASED SCHOOL ACCOUNTABILITY SYSTEMS

Test-based accountability systems often involve the threat of sanctions for teachers or school principals that do not meet specified standards. These systems tend to produce positive effects, but also often some unintended, negative effects. This study examines a school accountability program, which introduced substantial economic rewards targeting schools rather than individual teachers, without any sanctions attached. The design of the program combined with administrative data on all students and their parents in the country allow us to identify the effect of the program using both a regression-discontinuity design and a difference-in-differences design. Despite the large economic incentives, the results suggest that the program had only small positive effects on the targeted students, with no unintended consequences.

Multiple studies have examined school accountability systems that use sanctions or a combination of sanctions and rewards to motivate schools to enhance performance. Reviews of this literature tend to agree that these accountability systems produce (modest) positive effects—but also often some unintended, negative consequences such as prioritizing subgroups of students, reclassifying students into special education programs, or cheating on test results (Deming & Figlio, 2016; Figlio & Ladd, 2015; B. Jacob, 2017; National Research Council, 2011). One alternative approach to achieve the positive effects while avoiding negative spillovers could be to (i) use rewards without sanctions, and (ii) target schools rather than individual teachers. However, very few studies have rigorously examined school accountability systems with these policy design features.

In this study, we examine the effect of a recent policy in Denmark that offered support and large economic rewards to low-performing public schools. To earn the reward of around 200,000 USD, schools had to reduce the share of low-performing students across six exit exams in math, reading, and writing. With about 50 students at the targeted grade level, the rewards correspond to about 4,000 USD per student on average—which is most likely a lower-bound estimate of the actual incentive per targeted student given that some students were expected to score far above the

threshold for being classified as "low performing" even in the absence of the program.[1] In comparison, the early childhood education accountability system in North Carolina studied by Bassok and colleagues rewarded between 55 and 153 USD per subsidy-eligible student for achieving targets (Bassok et al., 2019, p. 846). The 200,000 USD in the program in Denmark correspond to about 4-5% of the yearly school budget. This is also substantially higher than the 1% reward in the Californian GPAP program studied by Bacolod and colleagues (2012). During the two-year program in Denmark, participating schools could earn the reward twice by reducing the share of low-performing students by 5 percentage points in the first year and 10 percentage points in the second year, relative to their baseline.

Two features of the policy design help us identify the causal effect of the program. First, to be invited for the program, schools' average percentage of low-performing students across three baseline years should be higher than a pre-determined threshold. This discontinuity in the percentage of low-performing students facilitates a regression discontinuity (RD) design to estimate the effect of the policy. Second, to ensure that the program was offered to schools in all parts of the country, region-specific thresholds were assigned to six geographical areas. As the region-specific thresholds vary considerably, we can use a difference-in-differences design (DiD) combined with coarsened exact matching (Iacus et al., 2012) to compare changes in percent low-performing students in invited and non-invited schools from different regions but with the same percentage of low-performing students in the baseline years. Whereas the DiD design relies on the stronger identifying assumption about common trends, it has the advantage of having more statistical power

[1] Low-performing students were defined by the policy as students that scored below a grade point average (GPA) of 4 (equivalent to D on the European ECTS grading scale) in both two math exit exams and four reading and writing exit exams.

than our RD estimation. By comparing results from both designs, we can assess the robustness of the findings.

Due to rich administrative data available for the full population of public school students, we can evaluate the impact of this policy not only on the targeted students in the targeted exams, but also for outcomes in other subjects, other groups of students at the schools, and alternative outcomes such as student well-being. The DiD estimates indicate that the program reduced the share of low-performing students with about 4 percentage points in the first year. The RD estimates are somewhat smaller (around 2 percentage points) and insignificant in most specifications. We do not find that effects of the program were larger in the second year, even though the schools had a year more to increase student achievement, and about 60% of the schools earned the reward after the first year (and thus had additional resources at their disposal). Finally, we find no indications that the program had negative effects on low-stakes subjects, on groups of students whose performance was not incentivized, or unintended impact on student well-being. Qualitative interviews and survey data from teachers and principals in both invited and non-invited schools support this conclusion. First, these supplementary data sources suggest that invited schools made some effort in raising students with exam grades expected to be close to the threshold. Second, the data also indicate that teachers within the invited schools felt strong opposition to improve achievements in the incentivized subjects for children close to the threshold at the cost of other students or other outcomes for the targeted students.

In sum, our results indicate that a reward-based accountability system targeting the school-level may avoid the potential negative consequences of stronger, sanction-based systems, but at the cost of only small positive effects. Despite the substantial economic incentives, the fact that they were targeted at the schools and not at individual teachers may have reduced their effect on teacher

behavior. Thus, school-level rewards may water down the incentives to the extent that they become ineffective.

Research on School-Accountability Systems

In a review of research on incentives and test-based accountability in education, the National Research Council (2011) identified different features that distinguish accountability systems. Two important factors are the use of sanctions vs. rewards and whether incentives are targeting schools or individual teachers.[2]

Sanctions or Rewards

School accountability systems may introduce two types of consequences: rewards and sanctions. Positive consequences may include increased resources, greater school level autonomy, or teacher bonuses. Negative consequences may be withdrawal of autonomy, school restructuring or closure of the school (Figlio & Ladd, 2015).[3]

[2] Other aspects of school accountability systems such as the scope of the performance measures may also influence their effects (National Research Council, 2011). Figlio and Ladd (2015, p. 197) note: "On the one hand, holding schools accountable for a small set of outcomes provides incentives for schools to narrow the scope of the education they provide. On the other hand, a broad set of outcomes is more difficult to measure reliably and may blur the focus of school and district personnel." Moreover, incentives could be targeting the students rather than the teachers (Hvidman and Sievertsen 2019).

[3] Hanushek and Raymond (2005) classified state accountability systems as either "report-card accountability" or "consequential accountability." The former type publishes performance information so that parents and other stakeholdes may "vote with their feet" and select schools with better performance – but without tying any specific sanctions or rewards to the outcomes.

Although the No Child Left Behind (NCLB) program included both sanctions and rewards, severe sanctions for persistently low-performing schools were predominant, and some states even threatened all low-performing schools with explicit sanctions (e.g., reconstitution) (Dee & Jacob, 2011). As the bulk of school accountability studies are done on NCLB in the US, most studies have examined accountability system with sanctions targeted the school level.

Studies of school-based sanction systems (which may also sometimes include a reward element) generally tend to find positive effects, even though effect sizes are modest (Dee & Jacob, 2011; B. A. Jacob, 2005; J. Lee & Reeves, 2012). The preferred estimate for both the pre-NCLB and NCLB periods by the National Research Council (2011) was 0.08 standard deviations of the test score distribution. Figlio and Rouse (2006), West and Peterson (2006), Rouse et al. (2013), and Chiang (2009) exploited differences in pressure within accountability systems because of lower performance ratings and found positive effects of up to 0.20 standard deviations, though most estimates were between 0.05 and 0.10.

At the same time, sanction-based school-accountability systems have been generating negative, unintended consequences ranging from mild to more severe. Some studies have found that the systems affected only some of the target groups of students or subjects. For instance, Jacob (2005) found positive trends in both math and reading scores on the high-stakes tests following the introduction of sanction-based accountability in the city of Chicago, whereas results on the lower-stakes state test were less pronounced. Macartney (2016) found that schools and teachers in North Carolina responded to value-added performance targets by reducing effort in baseline periods. Other studies have found more severe negative effects. Some schools have been shown to reclassify students into disability categories in order to pull them out of the statistics (Cullen & Reback, 2006; Deming et al., 2016; Figlio & Getzler, 2006; B. A. Jacob, 2005), use disciplinary procedures to suspend low-performing students from school when the tests are given (Figlio, 2006), and even

downright cheating by changing student answers on high-stakes exams (B. A. Jacob & Levitt, 2003).

From one perspective, rewards and sanctions are completely symmetrical: Rewarding high-performing schools with a bonus is similar to sanctioning low-performing schools by taking away the bonus they would otherwise have received. The difference between sanctions and rewards is a matter of how the baseline is framed. Yet, a substantial amount of research in psychology and behavioral economics has demonstrated the effects of negativity bias and loss aversion. People tend to put more emphasis on numbers presented negatively and are more unwilling to give away something they had owned than they are willing to gain something they had not owned (Kahneman et al., 1991; Kahneman & Tversky, 1979; Tversky & Kahneman, 1991). Rewarding with a payment framed as "bonus" may therefore be less effective than reducing payment, which would be framed as a fine or a sanction.[4]

Targeting Schools or Teachers

An emerging literature has studied reward-based teacher incentive systems using credible research designs. Using random variation in measurement error in the assignment to treatment variable, Lavy (2009) found that teachers rewarded with cash bonuses improved test taking rates, conditional pass rates, and mean test scores. In contrast, Fryer (2013) found no evidence that teacher incentives increased student performance, attendance, or graduation in a randomized trial. If anything, teacher incentives decreased student achievement. Sprinter et al. (2012) found no effect of

[4] In support of the notion that rewards and sanctions may not work symmetrically because of loss aversion, Fryer et al. (2012) showed that teachers paid in advance and asked to give back the money if their students did not improve sufficiently increases math test scores, whereas similar incentives in the standard, reward-based fashion resulted in smaller and statistically insignificant results.

bonuses awarded to teams of teachers. Dee and Wyckoff (2015) studied an accountability system from the District of Columbia that combined dismissal threats for low-performing teachers with economic bonuses for high-performing teachers. They found positive effects on the performance metric used for accountability, which included a combination of student performance data and classroom observations.

Targeting schools may, all else being equal, weaken the incentives for individual teachers, because it may not affect teachers' private incentives, or because it may induce free riding.[5] Some empirical evidence support the notion that targeting groups of teachers (or schools as organizations) rather than individuals weakens the incentives. Whereas Goodman and Turner (2013) found overall small—if any—effects of group-based teacher incentives, they found that schools for which incentives to free ride were weak experienced some increases in student achievement. Muralidharan and Sundararaman (2011) evaluated a randomized trial in India in which one treatment arm received teacher performance pay while a second treatment arm received school resources of a similar value. They found positive, but smaller effects of the group-based incentives than the individual incentives. Even though the National Research Council (2011, pp. 67, 83) notes that the high level of teacher absenteeism found in the study (25 percent) suggest that teacher incentives may work differently in the Indian case, the study by Muralidharan and Sundararaman (2011) is a strong indication that it matters whether incentives are connected to schools or individual teachers directly.

[5] Given that there are gains to cooperation or complementarities in production, group incentives could also yield better results than individual incentives (Hamilton et al., 2003; Itoh, 1991). Students' performance will often be affected by multiple teachers (and other staff at the school). Students may perform better in math if they have become better at reading and understanding the math problems and if they generally feel safe and secure at the school.

In sum, although extensive research has examined the effects of sanction-based school accountability systems—and some studies have examined reward-based teacher incentive systems—few studies have examined reward-based school accountability systems. Nevertheless, existing research may suggest that such systems provide weaker incentives and therefore are less likely to have both the positive effects and the negative spillovers found by previous studies.

Institutional background and program details

After ten years of compulsory education in a comprehensive school system, students in Denmark take the school-leaving examination before the transition to upper-secondary programs.[6] The school-leaving exams consist of a set of mandatory tests and a small number of elective tests.

Until 2014, the school-leaving examination was relativity low stake for students.[7] However, the stakes at the school-leaving examination have become higher for students. As of 2015, admission to the vocational upper secondary track has been contingent on passing the exams in the core subjects of (Danish) language and math. Moreover, as of 2019 admission to high-school programs has depended on performance at the school-leaving examinations.[8] Thus, the introduction of test-based

[6] There is one compulsory pre-school year plus 9 years of elementary/lower secondary education.

[7] Exams were low stake for the students, but not for schools. Publishing the schools' performance has been part of the (non-consequential) accountability system since 2002.

[8] While there is a range of different requirements, the core admission requirements are based on performance at the school-leaving examination. For vocational education, this means passing the exams (i.e. at least grade 02) in mathematics and language, and for high-school, this means attaining a mean grade of at least 5 in the four core subjects mathematics, Danish language, English and Science. The scores are reported on a 7-tiered grading scale that directly translates to the international ECTS scale. The scores in the grading scale are 12/A, 10/B, 7/C, 4/D, 02/E, 00/Fx, -03/F. Grade 02 is the minimum pass grade.

admission has raised the bar for entering upper-secondary programs particularly for academically weak students.[9]

At the school level, the publishing of performance tables of results from the school-leaving exams has been part of the non-consequential accountability program since 2002. The "Raising student achievement" program (RSA)[10], which is evaluated in this paper, is an additional component (targeting academically weak schools) to the universal accountability program.

The 'Raising Student Achievement' Program (RSA)

The RSA introduces strong incentives to improve school effectiveness in schools with high proportions of educationally struggling students by offering financial rewards to schools if they can reduce the proportion of students who score below some pre-designated score level (regarded as "adequate achievement") in language and mathematics.

The RSA program differs in some important ways from many other accountability programs. First, RSA introduces large financial rewards rather that sanctions. Second, the incentives are targeted the school level rather than teachers.

[9] While failing to pass the test-score threshold does not necessarily mean that they are precluded from upper-secondary programs, admission requires an extra effort as they would have to pass extra tests and interviews.

[10] In this paper, we have termed the program based on one of the names used by the Ministry of Education for the program Elevløft, meaning "Raising student (achievement)". In Denmark, this program is also known by the name "Skolepuljen".

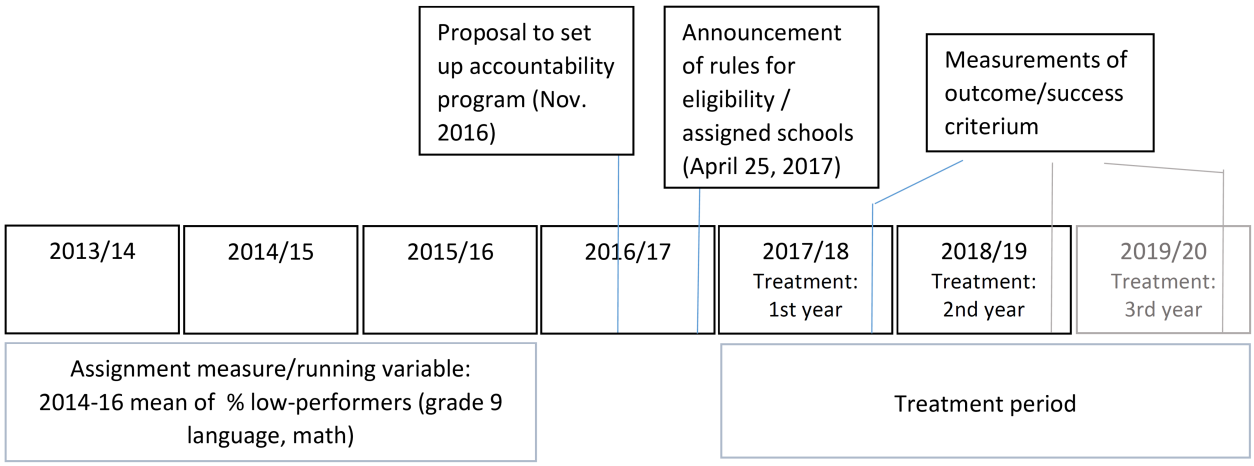


Figure 1. Timeline for school accountability program

Figure 1 shows the timeline for the program. The program is targeted schools with many "low-performing" students, which is defined as a student having a mean test score below grade 4 (corresponding to grade D on the international ECTS-scale) in mathematics or language (or both) on the final exam. The examination in language consists of tests in four subject areas and students get separate grades for each of them. Mathematics has two subject areas. The mean test score in language is the simple average of the four scores for language and the mean for mathematics is the simple average of the two math scores.

To be eligible for the program, schools had to fulfil two requirements. The first requirement is based on the schools' *share* of low-performing students over a pre-program three-year period (2014-16). The population share of low-performing students is 22% (2016, public school regular classes), but this share varies widely across schools (between 2% and 61%).[11] To ensure that the program targeted schools in all geographical areas of the country, the number of invited schools was proportional to the number of 9th grade students within six geographical regions (i.e., each of the

[11] To comply with rules of data protection, this is calculated on a trimmed sample excluding the 1% school cohorts at the extremes.

country's five administrative regions and the capital city, Copenhagen[12]). Thus, the number of schools invited varied between 7 in Copenhagen and 30 in Central Jutland. Within regions, the schools with the highest shares of low-performing students were invited to the program. Schools were ordered by their percentage of low-performing students within regions and—starting from the school with the highest share—schools were invited into the program until the number of schools allowed for each region was reached. As schools with high shares of low-performing students were not equally distributed across regions, the share of low-performing students at the school required for invitation varied considerably across these six geographical areas.

The second requirement for eligibility is based on the *number* of low-performing students. Schools must have more than an average of 11 low-performers in their 9th grade cohorts during the three-year baseline period of 2014-2016. This rule means that the program targets schools not only based on the share of low performers, but that there must be a 'critical mass' of low performers. The number criterion was constant across all regions in the country.

The RSA program rewards participating schools if they reduce the share of low-performing students by 5 percentage points in the first year and 10 percentage points in the second year relative to their baseline. A third year, requiring a 15-percentage points reduction in low performers, was originally announced, but after a shift in the government from right-wing to left-wing, the bonus program was terminated after two years.

The size of the rewards varied between 1.3-1.5 million DKK (corresponding to roughly 200,000 USD) depending on the school size. This is about 4,000 USD per student at the targeted 9th grade

[12] The Capital region has a high concentration of schools with large numbers of weak students in a single municipality (Copenhagen). To achieve a wider geographical spread of assigned schools across the Capital region, the municipality of Copenhagen was treated as a separate entity (and accorded a share of invited schools corresponding to its share of 9th graders in the region).

level, and about 4-5% of the yearly school budget. The monetary reward is given to the schools for unrestricted use. However, even though rewarded schools have considerable discretion in using these funds, the funds are unlikely to be used as personal rewards to teachers or school leaders, but are more likely to be used for extra teacher resources or school site purposes, such as instructional materials and equipment. Also, anecdotal evidence suggests that schools engaged private firms to help boost performance during the first year of the program. In some cases, this help was given on a 'no win, no fee' basis, meaning that some schools must use part of their reward to pay for these services.

In addition to the monetary incentive, the RSA program also provided guidance on how to boost performance of low achievers. This guidance included advice as to which evidence-based tools are considered effective to increase results for weak learners, counselling, and a forum to exchange ideas and experiences with other participating schools.

Data

We obtained a dataset from the Ministry of Education containing a list of all public schools that were invited to the program. This dataset also included information on the number and share of low-performing students in the three baseline years at each school as well as information on whether the schools accepted the invitation and participated in the RSA program.

We merged the school-level data with student-level data retrieved from administrative registers hosted by Statistics Denmark that is linked to the school-ID via a unique registration number. These data provide the full population of students and contain reliable information on test scores and students' family background, as well as grade-level identifiers.

The main outcomes are exam scores at the 9th grade school-leaving exams. The data include scores for each subject area. We standardize exam scores to a distribution with zero mean and a unit

standard deviation. To examine alternative outcomes, we add data on student well-being that are derived from a survey covering the entire population of 0-9 grade students. Well-being is assessed using three indicators validated by Andersen et al. (2020) as well as a general well-being measure.

For the main analysis, we use information for all ninth graders in regular public schools in the first and second year of the program. The dataset contains information on roughly 82,000 students in approximately 800 schools.

As part of a governmental evaluation of the program, surveys and interviews were conducted among teachers and principals at both participating and non-participating schools. We use this qualitative material to nuance our understanding from the statistical analyses.

Empirical strategy

The main challenge to estimating the causal effect of RSA on student achievement and other outcomes is that participation in the program is not randomly assigned. As assignment is based on previous performance, simple comparison of schools that participated in the RSA program with schools that were not invited would yield biased estimates.

To address the non-random assignment of schools to the RSA program, we exploit the unique features of the eligibility rules in the RSA to identify causal effects of the program. Figure 2 illustrates the different sources of variation provided by the RSA assignment rule. First, the use of cutoffs in the assignment rule provides a discontinuity in the probability of treatment assignment just around the cutoffs. Second, as the cutoff points vary across regions, there is variation in treatment assignment for schools with similar baseline performance but placed in different regions. As Figure 2 shows, the number criterion (more than 11 low-performing students on average) was the same in all six areas (the horizontal grey lines). However, the share criterion (the vertical grey lines) varied from about 25% in Central Jutland to about 50% in Copenhagen. Only schools that

fulfilled both criteria were invited. Yet, Figure 2 also shows that not all invited schools (schools in the north east corners above both grey cutoff-lines) accepted the invitation. For example, in Copenhagen none of the invited schools accepted the invitation. Overall, 104 schools participated. Of the 104 schools participating in the RSA, 88 also accepted to receive guidance on how to boost performance of low achievers.

In our analyses, we apply two empirical strategies that exploit the different sources of variation provided by the RSA: a regression discontinuity (RD) design and a difference-in-differences (DiD) design.

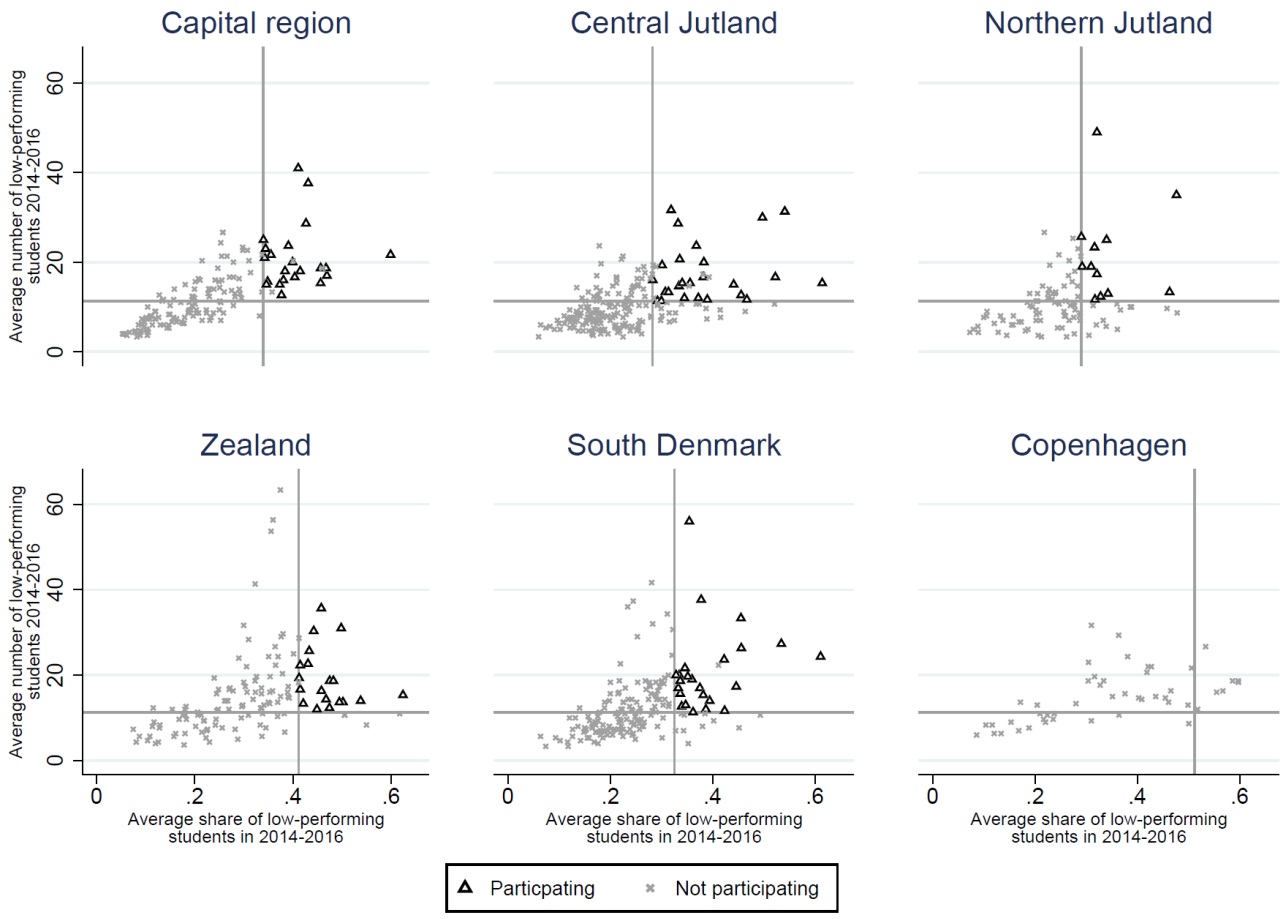


Figure 2. Invited and non-invited schools by five regions and Copenhagen based on region-specific cut-offs for average share of low-performing students in baseline years 2014-16, and common cutoff for more than 11 low-performing student in average across baseline years. Invited schools are in the upper right squares. Triangles designate participating schools.

Regression Discontinuity Design

We use a regression discontinuity design that compares schools that just barely were eligible with those that just barely were not eligible to isolate the causal effect of being invited to the program on our outcomes of interest. As running variable, we use the *share* of low-performing students at the school. As previously shown, there are different cutoffs for each region varying between 25-50% low performers in baseline years, which lead to a multiple cut-off RD design (Cattaneo et al., 2016). We normalize the running variable by centering it within each region and labeling the cut point 0, in order to pool the regions into a single analysis. Figure 3 illustrates the assignment mechanism for the share of low performers with the region-specific cutoffs pooled and normalized. We limit the RD estimation sample to schools meeting the criterion of having more than 11 academically low performing students on average in the three baseline years.[13] There is full compliance below the cutoff, i.e., no non-invited schools receive treatment. Above the cutoff, most but not all invited schools comply. Therefore, we estimate both the effect of being invited (ITT estimators), and the effect of actually participating in the program (LATE estimators). Note that since we have six regions with different cutoffs, we are not confined to estimate the effect at one specific cutoff but at six cut-offs across the distribution of the schools' percentage of low-performing students.

[13] The number of low-performers at baseline provides and alternative running variable (see Appendix A, Figure A1). As the assignment variables use different metrics, we follow Wong, Steiner & Cook (2013) and refrain from estimating a combined treatment effect across both frontiers. We focus on the frontier-specific effect for the *share* of low-performers assignment variable because this cutoff provides the highest number of observations and hence, we expect this cutoff to yield more precise estimates.

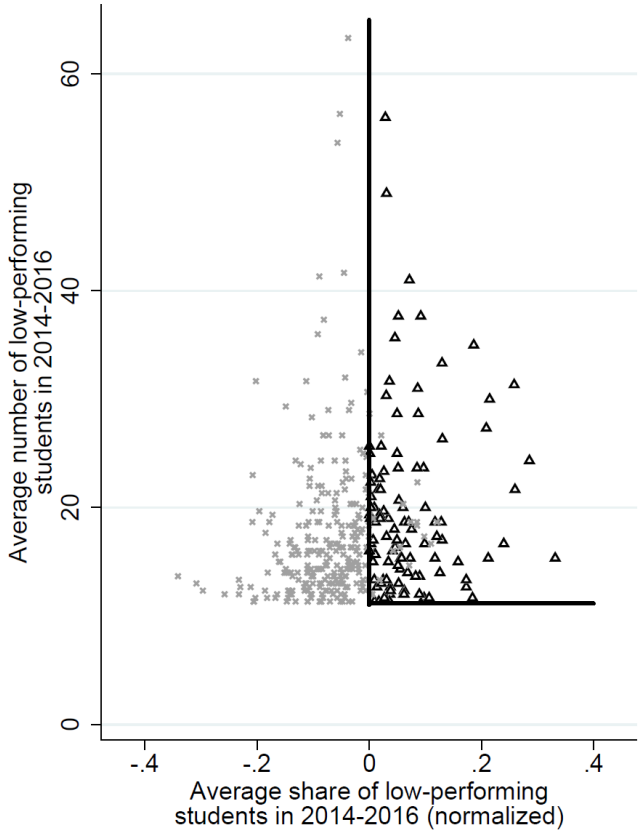


Figure 3. Invited and non-invited schools by average share of low-performing students. School that are not invited due to the absolute number criterion are excluded. Share of low-performance students is normalized across regions to have cut-off value set to 0. Triangles designate participating schools. The black line shows the frontier for assignment to the RSA program.

Invitation Effects (Reduced Form) and Participation Effects (LATE)

Formally, a school is invited to participate in the RSA program if the normalized running variable $s_s \geq 0$. We denote this by an indicator, d_s. If d_s=1, the school is invited for the program. If d_s=0, the school is not invited. The causal effect of attending a school that is invited to the program can be then estimated by β_1 in the following reduced-form specification:

$$outcome_{is} = \beta_0 + \beta_1 d_s + \beta_2(1 - d_s)f(s_s) + \beta_3 d_s f(s_s) + \beta_4 COInd_s + \beta_5 X_{is} + e_{is} \qquad (1)$$

In the main analysis, $outcome_{is}$ is 1 for students reaching the minimum competency threshold and 0 otherwise. $f(s_s)$ is a continuous function of the (normalized) distance of each school to the cutoff in the pooled data. We interact $f(s_s)$ with the invitation indicator d_s to allow for different slopes on each side of the cut-off. The key requirement for identification in a RD design is that we can separate the effects of the threshold (β_1) from the continuous function $f(s_s)$. Gelman and Imbens (2019) suggest that using global high-order polynomials of the continuous function could lead to misleading results. We therefore estimate the model using a non-parametric method: local linear regression (with a triangular kernel) and explore alternative bandwidths. We use a linear function as the main specification but test the robustness by using a second-order polynomial. Moreover, we include cut-off fixed effects ($COInd_s$) to control for the fact that observations in our pooled specification come from cut-offs that are spread widely across the distribution of the running variable (from 25 to 50% low performers). Last, to increase the precision of our estimates, we include a number of student characteristics (gender, ethnicity, age, pre-program reading and math scores, mother's education) represented by the vector X_{is}.

In the reduced-form specification above, the key independent variable d_s is the *invitation* to participate in the program (i.e. intent-to-treat effects), not *actual* participation. We also estimate the local average treatment effect for those schools induced to participate if invited (Hahn et al., 2001) by using two-stage least squares estimates, with the school's decision whether or not to accept the invitation to participate as the outcome in the first stage. This is the ''fuzzy'' variant of RD, in which intent-to-treat effect is estimated and scaled by the level of compliance with the threshold rule to obtain average causal effects for those receiving the treatment.

Validity of the RD model

In the empirical framework described above, β_1 will yield unbiased estimates of the causal effect of the RSA if there are no systematic differences between those who are just above and just below the cut-off (D. S. Lee & Lemieux, 2010).

To examine the plausibility of this assumption, we conduct a set of validation checks. First, we look at the density of the running variable around the cutoff and find no evidence of systematic manipulation of the running variable (Appendix B, Figure B1). Second, we examine whether, near the cutoff, treated units are similar to control units. These variables can be divided into two groups: variables that are determined before the treatment is assigned (predetermined covariates) and variables that are determined after the treatment is assigned but could not possibly have been affected by the treatment (placebo outcomes). We provide visual and regression evidence (by estimating RD regressions using covariates as left-hand side variables) and do not find statistically significant differences, which is reassuring (Appendix B, Figure B2, Tables B1 and B2).

Third, the design of any school accountability program is important for the incentives the program provides and potential unintended behavior, which could violate our identification strategy. We address different potential behavioral responses. (1) In the RSA, all students educated in mainstream classes count in the performance calculation, including students with special educational needs (as long as they attend mainstream classes). Students enrolled in special classes in mainstream schools, however, do not count. Although this incentivizes strategic behavior, we find no evidence that schools assign low performers to special classes in response to the program (Appendix C, Table C1). (2) Students who do not take all six required tests do not pass the threshold for 'acceptable achievement' and count as low performers. Thus, in contrast to other accountability systems, schools have an incentive to encourage students to sit the exam. Still, we

find no evidence that the program decreases the probability of a student missing an exam (Appendix C, Table C2)

Overall, the validity checks conducted do not provide evidence of a violation of the identifying assumption of the RD design.

Difference in differences design

As an additional strategy, we use a difference-in-differences design (DiD) that compares the development in outcomes between the schools participating in RSA and a control group that did not participate in the RSA. The underlying assumption is that participating schools in RSA and the control group would follow the same development in outcomes had there been no RSA program. One concern is that schools were assigned to the RSA based on their share of low-performing students being in top of the distribution within their region. Participating school are therefore even in the absence of the program expected to improve their performance compared to their baseline due to mean reversion.

However, the RSA assignment mechanism with regional-specific cutoffs makes the creation of a suitable control group possible. For values of the running variable above the lowest cutoff, there are schools in different regions with similar shares of low-performers, but with different invitation assignment due to the region-specific share-cutoffs.

Specifically, we apply coarsened exact matching (CEM) prior to our DID estimation to ensure that our analytical sample only include controls that are similar to the treated schools. Specifically, we match on the two eligibility criteria in the RSA, namely the number and the share of low-achieving students. CEM temporarily coarsens the data—i.e., divides the schools into strata based on their share and their number of low-performing students at baseline. The procedure then exact matches on the coarsened data, such that participating and non-participating schools with similar values on these two measures are matched (within strata).

The matching procedure employed divides our data into 50 strata.[14] Of these 50 strata, 19 contain both participating and non-participating schools. Figure 4 shows the matching procedure graphically. Schools belonging to a stratum are all marked in the same color. Participating schools that could be matched are marked by triangles and non-participating schools are marked by x's. Participating schools that could not be matched are marked by circles. Data from the 19 strata with matches are used in the outcome estimations, which implies that we only include non-participating schools and participating schools that are similar with respect to the two assignment variables. Appendix D, Table D1 provides an overview of the number of participating and non-participating schools that are trimmed from the DiD estimation.

[14] We provide robustness checks with varying numbers of strata in Appendix D, Table D3.

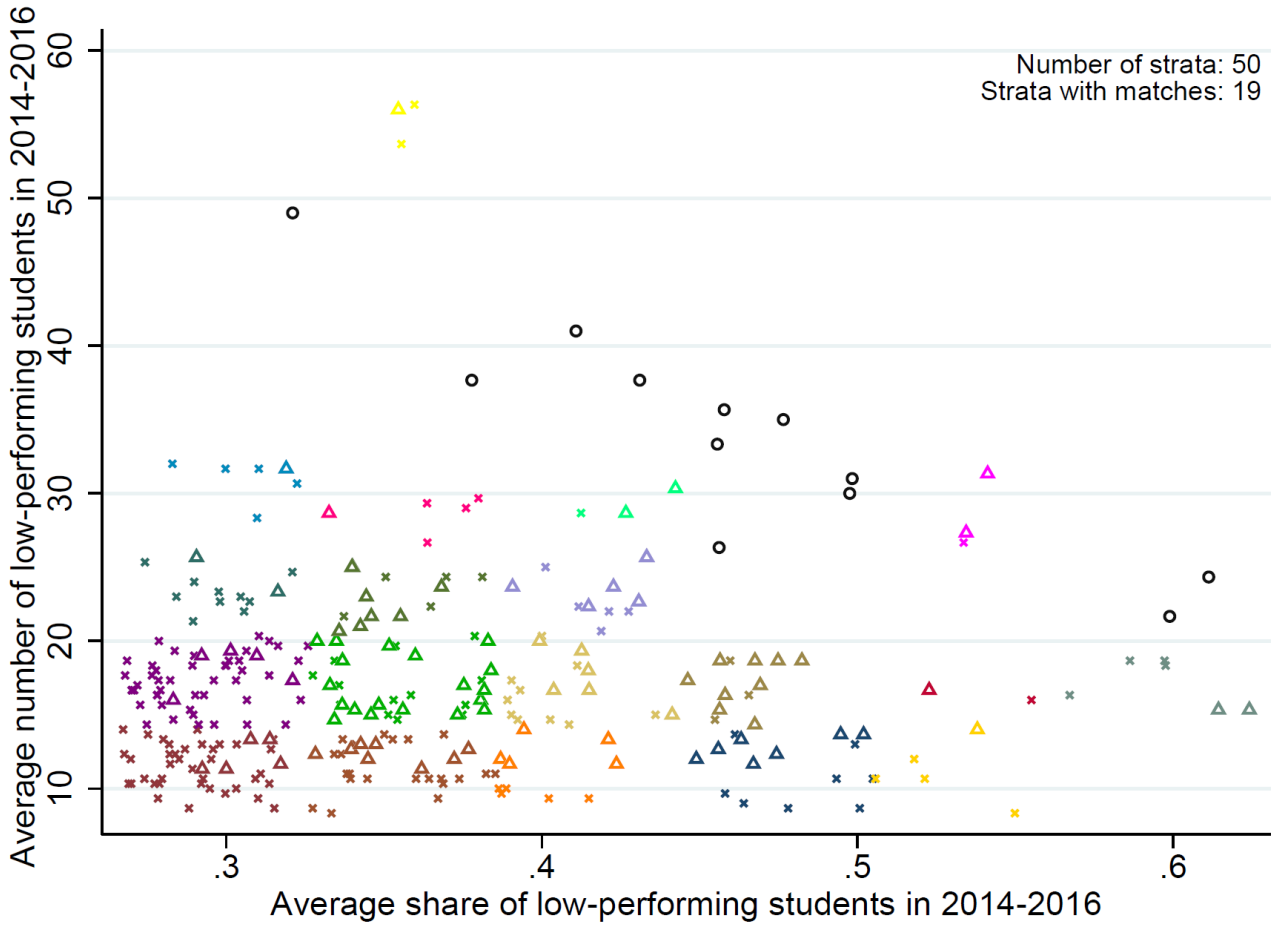


Figure 4. Strata matched using coarsened exact matching. Colors indicate strata. Triangles denotes participating schools, and X's denote non-participating schools. Circles denote schools that were not matched.

The matched DiD model

Formally, the effect of the program can be estimated by the following school-level regression:

$$Outcome_{st} = \gamma_0 + \gamma_1 T_{st} + \gamma_2 t_{st} + \gamma_3 (T_{st} * t_{st}) + \gamma_4 X_{st} + \epsilon_{st}, \quad (2)$$

where T_{st} indicates treatment for school s, and t indicates time period. γ_0 is the baseline of the control group, γ_1 is the initial difference between the control group and the treatment group, γ_2 is the common trend and γ_3 is the parameter of interest, the effect of participating in the RSA program.

To use maximal information, all schools in a stratum are used for matching, resulting in strata with unequal number of treated and control units. To compensate for the differential stratum sizes, we estimate the DiD model by weighting observations according to the stratum size (Iacus et al., 2012).

Validity of the DiD model

A way to evaluate the validity of the DiD approach is to estimate DiD regressions using covariates (aggregated to the school level) as left-hand side variables. Different trends in covariates between treated and controls would be a concern. Table D2 in the Appendix shows the results (only the coefficients of interest are shown). The point estimates indicate only small differences in the changes in the covariates between treated and control schools that are all insignificant. To adjust for small remaining imbalances, we include school-level shares or averages of student characteristics, represented by the vector $\boldsymbol{X}_{st}$. As we show in the results section, we also conduct supplementary analyses where we use outcomes for 2017 as a placebo outcome and find no evidence that would suggest that the common-trends assumption is violated.

In sum, both identification approaches have strengths and weaknesses. While the RD estimator is known to have strong internal validity, the DiD estimator has higher statistical power (than the RD design) because it uses information on a larger set of schools than those just around the eligibility cutoffs. We use the combined evidence from both methods to draw conclusions.

Descriptive statistics

Before we present the main results, we describe our sample and the selection into treatment. Table 1 provides descriptive statistics for non-invited schools, invited but not participating schools, and participating schools. As schools were selected based on their previous performance, it is not surprising that we find students in invited schools having lower test scores, lower parental

education, and are more likely to have an immigrant background than students in non-invited schools. Interestingly, students in invited, but not participating, schools were even more likely to have an immigrant background, lower test scores and lower maternal education than in schools that accepted the invitation. Thus, participating schools are not representative for all invited schools. These differences are primarily driven by schools in the capital, Copenhagen, where none of the invited schools participated in the RSA (see Figure 2). Since Copenhagen schools are 50% of the sample of non-participating schools that were invited, they heavily influence the summary statistics for this group.

Table 1. Student characteristics by schools' invitation and participation status

	Non-invited	Invited, not participating	Participating	Total
Immigrant	0.101	0.375	0.233	0.123
	(0.302)	(0.484)	(0.423)	(0.329)
Boy	0.512	0.488	0.511	0.511
	(0.500)	(0.500)	(0.500)	(0.500)
Age	15.67	15.74	15.72	15.68
(Jan 1, year of exam)	(0.403)	(0.490)	(0.449)	(0.411)
Age squared	245.8	247.8	247.4	246.0
	(12.75)	(15.64)	(14.28)	(13.03)
Math test score	0.0952	-0.249	-0.216	0.0506
	(0.939)	(0.965)	(0.939)	(0.946)
Read test score	0.103	-0.243	-0.223	0.0559
	(0.911)	(1.013)	(0.995)	(0.931)
Mother, high-educated	0.388	0.221	0.234	0.365
	(0.487)	(0.415)	(0.423)	(0.482)

Means (standard deviations in parentheses)

Results

The main objective of the RSA program was to improve student outcomes, with the specific goal of reducing the school-level percentages of low-performing students (defined as scoring below grade 4 in language and math exams). Descriptive statistics show that in the first year (2018), about 60% of the participating schools reached the 5 percentage points-target and received the reward. In the second year, about 52% reached the 10 percentage points-target. Although these descriptive outcomes are consistent with the RSA having a positive impact, these results may have occurred even in the absence of the program. We, therefore, proceed to study the effect of the RSA with the RD and DiD designs. The first question we ask is whether the program succeeded in raising more students above this threshold in the first year of the program, first at the award threshold and then at other points of the test score distribution. We estimate effects for all students and separately by subgroups delineated by ethnicity, gender, and socio-economic status (SES) and we also present results for student well-being. Last, we show main results for the second year of the program.

Effects at Program Target

In terms of failing to reach the threshold of grade 4 in both the math exam and the language exams, Figure 5 generally shows a positive relationship between the failing rate in the first treatment year and the percentage of failing students in the baseline years (the running variable). However, at the cut-off (s_s=0), there is a small dip in the failing rate, which may suggest that the probability to fail is lower for treated students at the cut-off.

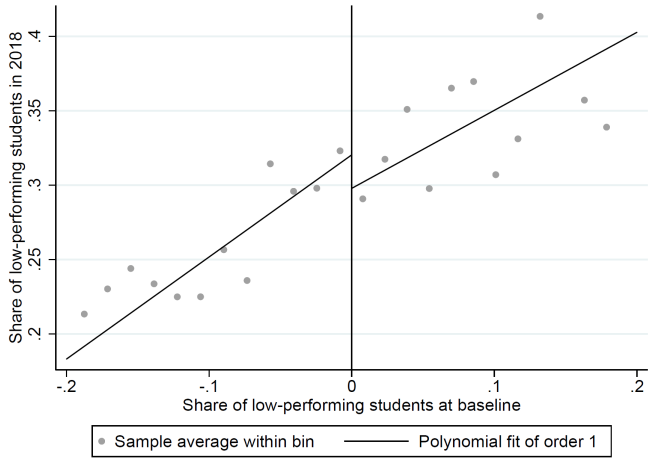


Figure 5. Change in share of low-performing students around the baseline cutoff (year 1).

We now turn to the formal results based on estimation of Eq. (1). Table 2 shows results from four different model specifications. All models use a linear specification and standard errors that are corrected for clustering at the school level. Columns (1) and (2) present the estimates from a sharp RD model, first without (1) and then with covariates (2). We also provide LATE results estimated by 2SLS. The first stage estimation is shown in Column (3) suggesting that passing the eligibility threshold increases the probability of participating in the program by about 87 percentage points (Appendix A, Figure A2, shows a graphical presentation of the first-stage effect). Results from the 2SLS estimation are shown in columns (4) and (5). As non-compliance at the discontinuity point is limited, the results are only slightly larger than the ITT estimates from the sharp model. The size and significance of the point estimates for the different model specifications in Table 2 vary slightly, but the sign is always negative suggesting that—on average—students profit somewhat from RSA.

Table 2. Main results for RD model. Effect on share of low-performing students (year 1).

	Sharp		First stage	Fuzzy	
	(1)	(2)	(3)	(4)	(5)
RD_Estimate	-0.03	-0.02	0.87**	-0.04	-0.03
	(0.02)	(0.02)	(0.06)	(0.02)	(0.02)
Polynomial	1	1	1	1	1
Bandwidth	.20	.20	0.20	.20	.20
Covar_Cutoff		X			X
n_eff	20506	20506	20506	20506	20506
N	21888	21888	21888	21888	21888

Clustered standard errors in parentheses + p<0.10, * p<0.05, ** p<0.01

Table 3. Robustness for sharp RD model. Effect on share of low-performing students (year 1).

	(1)	(2)	(3)	(4)	(5)	(6)
RD_Estimate	-0.04	-0.03	-0.02	-0.01	0.00	0.01
	(0.03)	(0.03)	(0.03)	(0.02)	(0.03)	(0.03)
Polynomial	2	1	2	2	1	2
Bandwidth	.20	Datadriven	Datadriven	.20	Datadriven	Datadriven
Covar_Cutoff				X	X	X
n_eff	20506	10587	15828	20506	8722	12731
N	21888	21888	21888	21888	21888	21888

Clustered standard errors in parentheses + p<0.10, * p<0.05, ** p<0.01

In Table 3, we present results from different specifications for the sharp RD model to assess the robustness of our main results. The specifications vary along three dimensions: whether they include control variables or not, include linear or quadratic functions, use fixed or data-driven bandwidth. The estimates from models without controls have negative signs throughout (columns (1)-(3)) and sizes are in the same range as in Table 2. The results with controls are all close to zero (columns (4)-(6)).

As a robustness check, we implement a placebo tests by estimating treatment effects using alternative values of the threshold (where there should not be any effect). We examine two fake cut-offs to the left and two to the right of the actual cut-off (i.e. schools with 5 and 10 pp. fewer/more low-performers than at the actual threshold). We estimate regressions for the main model

specification (column (2) in Table 2) using the fake-cutoffs and found all but one of the estimates to be insignificant. Results are graphically displayed in Figure 6.

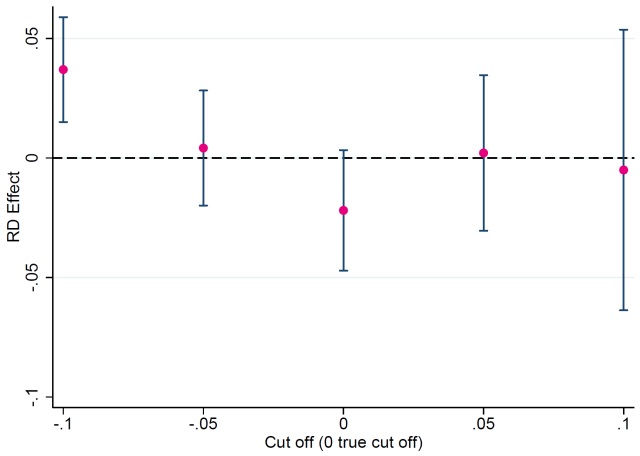


Figure 6. Change in share of low-performing students around the baseline cutoff (year 1).

Note: Estimations based on linear regression, bandwidth of .20, covariates and cutoff fixed effects included in sharp RD.

DiD results

In this subsection, we present results from the second identification approach that exploits variation in treatment across regions using DiD methods. Table 4 shows DiD results using automated coarsening in the matching procedure for different model specifications for 2018 (year 1). All estimations include indicators for year 2017 and 2018, an indicator for treatment and their interactions. The row titled Participants 2018 in Table 4 contains the main effect estimates (i.e. the interaction between the treatment and year 2018 indicators).

Column (1) presents results from a model without covariates on the unmatched sample, while column (2) restricts the estimation to the matched sample. Columns (3) and (4) separately add controls and school-specific time trends, and column (5) presents results for the full model. Results using different coarsening choices are shown in Appendix D, Table D3.

Table 4. Effect on share of low-performing students (2018, year 1). DiD models

	(1) Raw	(2) Matched	(3) Matched and covariates	(4) Matched and trend	(5) Matched, covariates and trend
2017	0.01**	-0.02	0.01	0.01	0.02+
	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)
2018	0.00	-0.02+	-0.01	0.03**	0.03**
	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)
Treatment	0.17**	0.00	0.01+	0.01+	0.01+
	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)
Participants 2017	-0.03**	0.00	-0.02	0.01	-0.00
	(0.01)	(0.02)	(0.01)	(0.02)	(0.01)
Participants 2018 (estimate of main interest)	-0.06**	-0.04**	-0.05**	-0.03**	-0.03**
	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)
Constant	0.23**	0.39**	-0.84	-0.02**	-1.08*
	(0.00)	(0.01)	(0.59)	(0.01)	(0.45)
Covariates			x		x
Trend				x	x
Observations	4065	1385	1385	1385	1385

Clustered standard errors in parentheses $^{+}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$

For all model specifications, the sign of the main effect estimate is negative and significant at the 1% level suggesting that students in participating schools on average do better than in other schools. The DiD results back our conclusion from the RD analysis that the RSA appears to lift more students over the minimum competency threshold. Overall, the RD and DiD results convey a picture that – whatever method and specification we use – the RSA seems to improve student outcomes, albeit the effect size in the full model is not large.

The common trends assumption is key for the validity of the DiD approach, positing that the average change in the comparison group represents the counterfactual change in the treatment group if there were no treatment. Equality of pre-treatment trends may lend confidence but this cannot directly test the identifying assumption which by construction is untestable.

In our set-up, we can use the outcome measure of 2017 as a pre-treatment outcome to provide some evidence on the common trends assumption. The row titled Participants 2017 in Table 4 contains the estimate for the interaction between the treatment and year 2017 indicator. The estimates for 2017 are all insignificant (except, as expected, for the raw, unmatched model 1) indicating that exam results at invited and non-invited schools followed a common trend before the onset of RSA, which is reassuring.

Subsample results

This subsection shows results by ethnicity, gender, and SES. Table 5 presents results using the RD model.[15] The signs of the estimates are all negative indicating that all subgroups profit from the RSA. The differences by subgroups are small throughout and probably not significant, suggesting that the RSA does not seem to affect these subgroups differently.

[15] The DiD analyses use aggregate (school-level) data, which is why we do not present subgroup results based on the DiD design.

Table 5. Effect on share of low-performing students (2018, year 1) for different subgroups. SHARP RD

	(1) Danish background	(2) Immigrant background	(3) Girl	(4) Boy	(5) Mother, low-educated	(6) Mother, high-educated
RD_Estimate	-0.02	-0.03	-0.01	-0.04+	-0.02	-0.02
	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
Polynomium	1	1	1	1	1	1
Bandwith	.20	.20	.20	.20	.20	.20
Covar_Cutoff	X	X	X	X	X	X
n_eff	16955	3551	9972	10534	13768	6738
N	18047	3841	10650	11238	14562	7326

Effects at other points on the grading scale

Above we examined whether RSA succeeded in lifting more students over the grade 4 threshold. An alternative question is whether there is a broader-based improvement in teaching. The RSA assessment measure is based on the share of students who achieve a certain proficiency target rather than being based on the average performance of students in a school or the value-added gains that students made from their scores in the previous year. When schools are assessed based on proficiency targets, they have a strong incentive to focus on students who are near the threshold, or on other students more likely to count for accountability (see also Figlio & Ladd, 2015; Neal & Schanzenbach, 2010; Reback, 2008).

We therefore examine whether the incentives in the RSA induced the participating schools to focus more on students expected to score just around the performance target and whether this has come at the expense of other points of the achievement distribution. In the Danish context, for example, a crucial threshold for admission to vocational upper secondary programs is passing the

language and math exam (i.e. at least grade 2). Below, we therefore test whether RSA affects students at other points of the test score distribution.

While the results in Tables 2-4 examine improvements at the performance target, Table 6 and Table 7 present results at other points of the test score distribution. These estimates are generally smaller than at the performance target (or even positive). These results suggest that the improvement in scores (if any) is located mainly at the RSA performance target, but is generally not detectable at other points of the test score distribution.

Table 6. Effect on other points in the grading scale (2018, year 1). Sharp RD

	(1) Score below 2	(2) Score below 7	(3) Score below 10
RD_Estimate	-0.01 (0.01)	-0.00 (0.02)	-0.00 (0.01)
Polynomium	1	1	1
Bandwith	.20	.20	.20
Covar_Cutoff	X	X	X
n_eff	20506	20506	20506
N	21888	21888	21888

Table 7. Effect on other points in the grading scale in year 1 (2018). DiD

	(1) Score below 2	(2) Score below 7	(3) Score below 10
2017	0.03	-0.04	-0.05^{+}
	(0.05)	(0.05)	(0.03)
2018	-0.03	0.09^{+}	0.02^{+}
	(0.03)	(0.05)	(0.01)
Treatment	0.03	-0.02	-0.02^{+}
	(0.03)	(0.04)	(0.01)
Participants 2017	0.06	0.11	0.05
	(0.07)	(0.08)	(0.04)
Participants 2018 (estimate of main interest)	-0.02	0.01	0.02
	(0.05)	(0.07)	(0.02)
Constant	-1.71	1.88	2.75**
	(1.99)	(2.60)	(0.91)
Observations	1385	1385	1385

All models using matched sample. Clustered standard errors in parentheses
Covariatates and trend variable included but not shown
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Well-being

Although the RSA only uses test scores as a performance metric for the award calculations, the program has the potential to—unintentionally—affect other outcomes such as student well-being. School accountability systems put extra pressure not only on school leaders and teachers, but possibly also on students. Thus, an unintentional effect of the RSA may be to lower student well-being.

To examine how the additional strain might affect the well-being of students, we use student level data for grade 9 students from a full population survey on student well-being that is conducted once a year by the Ministry of Education. We use eight items from the survey that have been validated to measure three socio-emotional skills (or personality domains): conscientiousness,

agreeableness, and emotional stability (Andersen et al., 2020). We supplement these three validated measures with a measure of students general well-being in school based on two items: ”Do you like your school?” and ”Do you like your class?”. Note that the scale is different from the scale for the main outcome and that the size of the estimates cannot be directly compared. Table 8 and Table 9 show the results. The RD estimates in Table 8 are negative except for agreeableness and all are insignificant. DiD estimates in Table 9 are positive, suggesting that the RSA harms well-being. However, none of the estimates is close to being significant.

Table 8. Effect on wellbeing (2018, year 1). Sharp RD

	(1) Wellbeing	(2) Conscientiousness	(3) Agreeableness	(4) Emotional Stability
RD_Estimate	-0.06	-0.04	0.03	-0.03
	(0.06)	(0.04)	(0.04)	(0.04)
Polynomium	1	1	1	1
Bandwith	.20	.20	.20	.20
Covar_Cutoff	X	X	X	X
n_eff	14564	14439	14576	14098
N	15503	15356	15514	15002

Table 9. Effect on wellbeing (2018, year 1). DiD

	(1) Wellbeing	(2) Conscientious-ness	(3) Agreeablenes s	(4) Emotional Stability
2017	-0.06	-0.10	-0.02	-0.06
	(0.15)	(0.10)	(0.10)	(0.11)
2018	-0.41**	-0.32*	-0.06	-0.30
	(0.15)	(0.14)	(0.13)	(0.19)
Treatment	-0.07	-0.08	-0.07	-0.13
	(0.10)	(0.10)	(0.11)	(0.09)
Participants 2017	0.10	0.15	-0.03	0.11
	(0.20)	(0.15)	(0.15)	(0.17)
Participants 2018 (estimate of main interest)	0.28	0.13	0.01	0.25
	(0.20)	(0.19)	(0.18)	(0.22)
Constant	-1.16	10.00+	7.83	9.06
	(7.23)	(5.69)	(5.29)	(7.07)
Observations	871	861	866	845

All models using matched sample. Clustered standard errors in parentheses
Covariatates and trend variable included but not shown
+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Year 2 results

Since 60% of participating schools received the year 1 reward, a substantial number of schools have had additional resources at their disposal in year 2. One might therefore expect RSA effects in the second year of the program to exceed those in year 1. However, the results for the second year (Table 10) are similar to the year 1 results. The RD models estimates effect sizes to between 1 and 2 percentage points and statistically insignificant, whereas the DiD models have more precision and find the effects to be between 3 and 4 percentage points.

Table 10. Effect on share of low-performing students in year 2 (2019).

	RD model		DiD model	
	(1)	(2)	(3)	(4)
Estimate	-0.02	-0.01	-0.04*	-0.03**
	(0.02)	(0.02)	(0.01)	(0.01)
Covariates		X		X
Cutoff-FE		X		
Trend				X
n_eff	21220	21220		
N	22676	22676	1385	1385

Clustered standard errors in parentheses
+ p<0.10, * p<0.05, ** p<0.01
Note: The RD models (sharp) are linear and with bandwidth 0.20.

DISCUSSION

This study evaluated the effects of a school accountability program that used only rewards—without any sanctions—and targeted schools rather than individual teachers or principals. This accountability design could potentially reap the benefits of sanction-based programs while avoiding the unintended effects of teachers feeling pressured to game or cheat the system. Nevertheless, very few existing studies have examined this type of accountability designs.

We find relatively small effects of the program. Point estimates vary a little depending on model specifications, but generally are around 2-5 percentage points, which means that in two classrooms of 25 students each, one or two more students make it above the target (an average grade of 4 corresponding to a D on the ECTS grading scale) due to the program. The effect estimates tend to be somewhat higher and more precisely estimated in the DiD models than in the RD models. The combination of the two identification strategies and the similarity in the results increases confidence that the true effect size is within the range of these estimates.

We found no unintended effects on other outcomes, even though we tested for several potentially negative side effects. First, we find no evidence that schools referred students to special needs classes to avoid that they negatively affected the schools' results. Second, the program did

not seem to affect the likelihood that students took the exams that were part of the incentivized performance score. Third, we find no evidence that schools focused their effort of students around the cut-off at the cost of students at other parts of the ability distribution. Fourth, we did not find indications that any increased performance pressure had effects on students' well-being.

An important question is why we do not see stronger effects of the program than we do. Interviews and survey data on both participating and non-participating schools (among both teachers and school principals) indicate that the program to some extend made schools focus more on the targeted group of low-performing students and their performance at the exams. So schools may have tried to pick some low-hanging fruit, but did not find measures that could lift this group of students more broadly. One reason for the relatively small effects may be that the financial incentives in the program were too small. However, compared to other school accountability programs with incentives at the school level, the rewards in this program was several times larger (cf. Bacolod et al., 2012; Bassok et al., 2019).

The fact that the rewards were targeted schools and not teachers' private incentives may be important. This is what might protect against negative unintended side-effects, but it may also mean that teachers did not have strong enough incentives to react to the program. A survey among the school principals shows that those who received the reward after the first year used the money for buying new equipment, in-service training of teachers, temporary employment of new teachers, field trips for students and several other activities. Whereas most of these activities may benefit the teachers, they had no direct impact on their private, economic incentives. In that sense the school-targeted incentives may have reduced both positive and negative effects.

The school-targeted incentives may also have increased the role of social, professional norms among the teachers. Since this was a collective enterprise, teachers may also have asserted norms saying that pursuing the goal of the reward should not be at the expense of other students, other

subjects than those that were targeted, let alone the wellbeing of the students. The strength of such professional norms and values is indicated by the fact that some schools, including all invited schools in the capital, Copenhagen, declined the invitation to participate in the program. Even if the schools did not want to change their behavior because of the program, they could have accepted the invitation and regarded it as a lottery that might pay-out even without any additional effort on their behalf. Indeed, our data from the intermediate year 2017 (after the baseline years, but before the first year of the program), indicate that around 40% of the schools received the reward of around 200,000 USD due to ordinary fluctuations in student performance across cohorts within the same schools. The fact that a substantial number of schools declined to participate therefore indicates some collective norms against this type of accountability system.

It is also worth noting that another constraint on the effects may be that schools already use most of the best-available methods to support the low-performing students. Rigorously evaluated, randomized intervention studies in education tend to find that new, theoretically promising interventions have rather small effects on top of treatment-as-usual in the control groups (e.g. Lortie-Forgues & Inglis, 2019). Classic theory of economic incentives assumes that with strong enough incentives, schools will find the solutions needed to get the rewards. Yet, if the best available methods are already employed, schools may not be able to react to the incentives in a way that improves performance further. This consideration is supported by the results showing that in the second year of the program—when schools had had two years to adjust their behavior to the new incentives, the requirements for the reward had risen from a target of 5 percentage points to 10 percentage points improvement, and when 60% of them had received the reward of about 200,000 USD after the first year—the effects were not larger than after the first year. The estimated effects were in the range of 1-4 percentage points.

These considerations regarding both professional norms shared by the teachers and the standard of the treatment-as-usual relate to an external validity issue. The present study examines low-performing schools in Denmark. And even though the large variation in region-specific cut-off values that were used to decide whether schools were invited to the program or not gives a broader basis for generalization than in typical RD studies with a single cut-off, the present study is limited the context of Denmark. Whether (low-performing) schools in other countries would have a different standard of treatment-as-usual or different professional norms that might influence the results, would need to be studied in future research.

Another study limitation is that we examine effects only immediately after one and two years. Long-term effects may differ from these immediate effects if schools have learned from the program or if the program gradually affects norms and behavior at the schools. The present program was terminated after a left-wing replaced the right-wing government that initiated the program, but due to the rich administrative data collected each year, it will be possible to follow up on the performance of the schools in the following years. Without any long-term effects, the costs of the present program seem to outweigh the rather modest positive effects on student performance—even if the program succeed in avoiding any negative effects on alternative outcomes.

REFERENCES

Andersen, S. C., Gensowski, M., Ludeke, S. G., & John, O. P. (2020). A stable relationship between personality and academic performance from childhood through adolescence. An original study and replication in hundred-thousand-person samples. *Journal of Personality*, *n/a*(n/a). https://doi.org/10.1111/jopy.12538

Bacolod, M., DiNardo, J., & Jacobson, M. (2012). Beyond Incentives: Do Schools Use Accountability Rewards Productively? *Journal of Business & Economic Statistics*, *30*(1), 149–163. https://doi.org/10.1080/07350015.2012.637868

Bassok, D., Dee, T. S., & Latham, S. (2019). The Effects of Accountability Incentives in Early Childhood Education. *Journal of Policy Analysis and Management*, *38*(4), 838–866. https://doi.org/10.1002/pam.22149

Cattaneo, M. D., Keele, L., Titiunik, R., & Vazquez-Bare, G. (2016). Interpreting Regression Discontinuity Designs with Multiple Cutoffs. *The Journal of Politics*, *78*(4), 1229–1248. https://doi.org/10.1086/686802

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics*, *93*(9), 1045–1057. https://doi.org/10.1016/j.jpubeco.2009.06.002

Cullen, J. B., & Reback, R. (2006). Tinkering Toward Accolades: School Gaming under a Performance Accountability System. In T. J. Gronberg & D. W. Jansen (Eds.), *Improving School Accountability* (Vol. 14, pp. 1–34). Emerald Group Publishing Limited. https://doi.org/10.1016/S0278-0984(06)14001-8

Dee, T. S., & Jacob, B. (2011). The impact of no Child Left Behind on student achievement. *Journal of Policy Analysis and Management*, *30*(3), 418–446. https://doi.org/10.1002/pam.20586

Dee, T. S., & Wyckoff, J. (2015). Incentives, Selection, and Teacher Performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, *34*(2), 267–297. https://doi.org/10.1002/pam.21818

Deming, D. J., Cohodes, S., Jennings, J., & Jencks, C. (2016). School Accountability, Postsecondary Attainment, and Earnings. *The Review of Economics and Statistics*, *98*(5), 848–862. https://doi.org/10.1162/REST_a_00598

Deming, D. J., & Figlio, D. (2016). Accountability in US Education: Applying Lessons from K-12 Experience to Higher Education. *Journal of Economic Perspectives*, *30*(3), 33–56. https://doi.org/10.1257/jep.30.3.33

Figlio, D. N. (2006). Testing, crime and punishment. *Journal of Public Economics*, *90*(4), 837–851. https://doi.org/10.1016/j.jpubeco.2005.01.003

Figlio, D. N., & Getzler, L. S. (2006). Accountability, Ability and Disability: Gaming the System? In T. J. Gronberg & D. W. Jansen (Eds.), *Improving School Accountability* (Vol. 14, pp. 35–49). Emerald Group Publishing Limited. https://doi.org/10.1016/S0278-0984(06)14002-X

Figlio, D. N., & Ladd, H. F. (2015). School Accountability and Student Achievement. In H. F. Ladd & M. E. Goertz (Eds.), *Handbook of Research in Education Finance and Policy* (pp. 194–210). Routledge. https://doi.org/10.4324/9780203788684-19

Figlio, D. N., & Rouse, C. E. (2006). Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics*, *90*(1), 239–255. https://doi.org/10.1016/j.jpubeco.2005.08.005

Fryer, J., Roland G., Levitt, S. D., List, J., & Sadoff, S. (2012). *Enhancing the Efficacy of Teacher Incentives through Loss Aversion: A Field Experiment* (Working Paper No. 18237; Working Paper Series). National Bureau of Economic Research. https://doi.org/10.3386/w18237

Fryer, R. G. (2013). Teacher Incentives and Student Achievement: Evidence from New York City Public Schools. *Journal of Labor Economics*, *31*(2), 373–407. https://doi.org/10.1086/667757

Gelman, A., & Imbens, G. (2019). Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs. *Journal of Business & Economic Statistics*, *37*(3), 447–456. https://doi.org/10.1080/07350015.2017.1366909

Goodman, S. F., & Turner, L. J. (2013). The Design of Teacher Incentive Pay and Educational Outcomes: Evidence from the New York City Bonus Program. *Journal of Labor Economics*, *31*(2), 409–420. https://doi.org/10.1086/668676

Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. *Econometrica*, *69*(1), 201–209. JSTOR.

Hamilton, B. H., Nickerson, J. A., & Owan, H. (2003). Team Incentives and Worker Heterogeneity: An Empirical Analysis of the Impact of Teams on Productivity and Participation. *Journal of Political Economy*, *111*(3), 465–497. JSTOR. https://doi.org/10.1086/374182

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, *24*(2), 297–327. https://doi.org/10.1002/pam.20091

Iacus, S. M., King, G., & Porro, G. (2012). Causal Inference without Balance Checking: Coarsened Exact Matching. *Political Analysis*, *20*(1), 1–24. https://doi.org/10.1093/pan/mpr013

Itoh, H. (1991). Incentives to Help in Multi-Agent Situations. *Econometrica*, *59*(3), 611–636. JSTOR. https://doi.org/10.2307/2938221

Jacob, B. (2017). The Changing Federal Role in School Accountability. *Journal of Policy Analysis and Management*, *36*(2), 469–477. https://doi.org/10.1002/pam.21975

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, *89*(5), 761–796. https://doi.org/10.1016/j.jpubeco.2004.08.004

Jacob, B. A., & Levitt, S. D. (2003). Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. *The Quarterly Journal of Economics*, *118*(3), 843–877. https://doi.org/10.1162/00335530360698441

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1991). Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias. *Journal of Economic Perspectives*, *5*(1), 193–206. https://doi.org/10.1257/jep.5.1.193

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263–291. JSTOR. https://doi.org/10.2307/1914185

Lavy, V. (2009). Performance Pay and Teachers' Effort, Productivity, and Grading Ethics. *American Economic Review*, *99*(5), 1979–2011. https://doi.org/10.1257/aer.99.5.1979

Lee, D. S., & Lemieux, T. (2010). Regression Discontinuity Designs in Economics. *Journal of Economic Literature*, *48*(2), 281–355. https://doi.org/10.1257/jel.48.2.281

Lee, J., & Reeves, T. (2012). Revisiting the Impact of NCLB High-Stakes School Accountability, Capacity, and Resources: State NAEP 1990–2009 Reading and Math Achievement Gaps and Trends. *Educational Evaluation and Policy Analysis*, *34*(2), 209–231. https://doi.org/10.3102/0162373711431604

Lortie-Forgues, H., & Inglis, M. (2019). Rigorous Large-Scale Educational RCTs Are Often Uninformative: Should We Be Concerned? *Educational Researcher*, *48*(3), 158–166. https://doi.org/10.3102/0013189X19832850

Macartney, H. (2016). The Dynamic Effects of Educational Accountability. *Journal of Labor Economics*, *34*(1), 1–28. https://doi.org/10.1086/682333

Muralidharan, K., & Sundararaman, V. (2011). Teacher Performance Pay: Experimental Evidence from India. *Journal of Political Economy*, *119*(1), 39–77. https://doi.org/10.1086/659655

National Research Council. (2011). *Incentives and Test-Based Accountability in Education*. National Academies Press. https://doi.org/10.17226/12521

Neal, D., & Schanzenbach, D. W. (2010). Left Behind by Design: Proficiency Counts and Test-Based Accountability. *The Review of Economics and Statistics*, *92*(2), 263–283. https://doi.org/10.1162/rest.2010.12318

Reback, R. (2008). Teaching to the rating: School accountability and the distribution of student achievement. *Journal of Public Economics*, *92*(5), 1394–1415. https://doi.org/10.1016/j.jpubeco.2007.05.003

Rouse, C. E., Hannaway, J., Goldhaber, D., & Figlio, D. (2013). Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure. *American Economic Journal: Economic Policy*, *5*(2), 251–281. https://doi.org/10.1257/pol.5.2.251

Springer, M. G., Pane, J. F., Le, V.-N., McCaffrey, D. F., Burns, S. F., Hamilton, L. S., & Stecher, B. (2012). Team Pay for Performance: Experimental Evidence From the Round Rock Pilot Project on Team Incentives. *Educational Evaluation and Policy Analysis*. https://doi.org/10.3102/0162373712439094

Tversky, A., & Kahneman, D. (1991). Loss Aversion in Riskless Choice: A Reference-Dependent Model. *The Quarterly Journal of Economics*, *106*(4), 1039–1061.

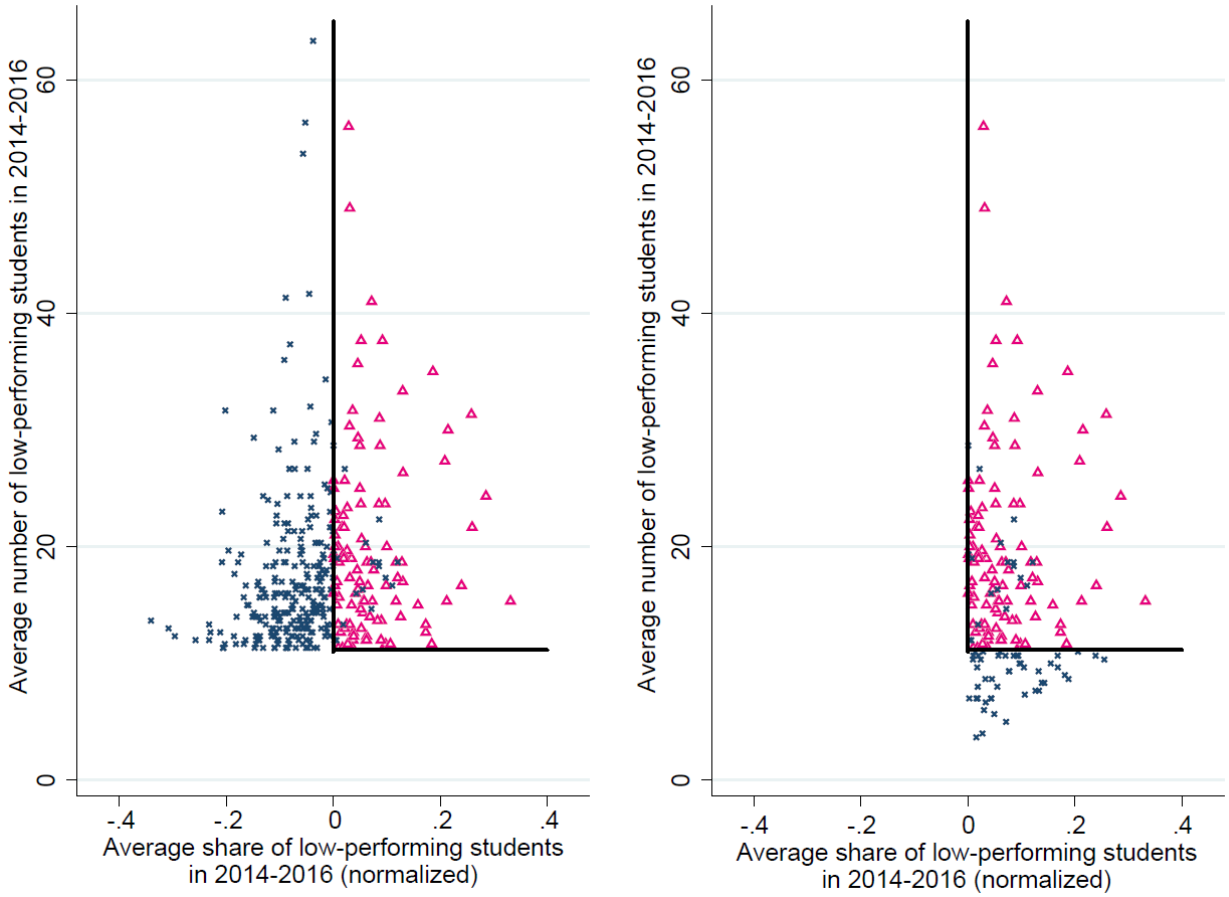
West, M. R., & Peterson, P. E. (2006). The Efficacy of Choice Threats Within School Accountability Systems: Results from Legislatively Induced Experiments. *The Economic Journal*, *116*(510), C46–C62. https://doi.org/10.1111/j.1468-0297.2006.01075.x

Appendix to

EFFECTS OF REWARD-BASED SCHOOL ACCOUNTABILITY SYSTEMS

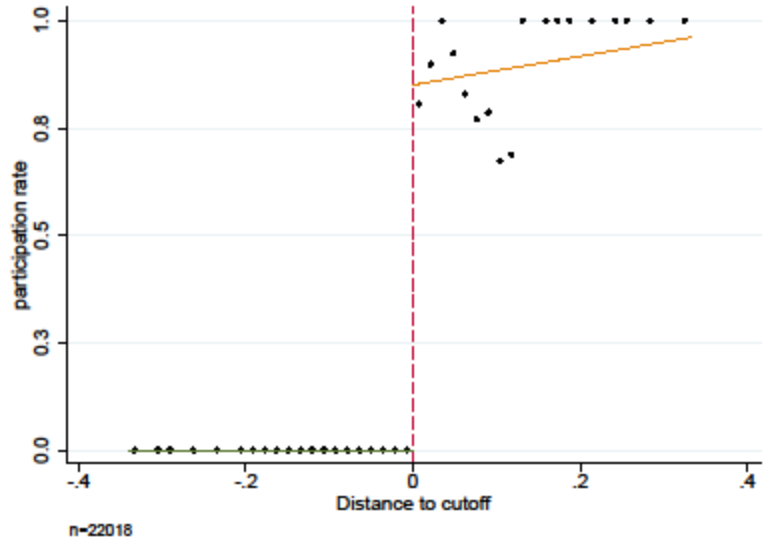
APPENDIX A: Eligibility and Participation in RSA

Figure A1. RD design, two cut-offs



Note: Each dot represents a school. Triangles are participating schools; squares are non-participating schools. The figure on the left illustrates the analytical sample created by the number criterion, where only schools with more than 11 low-performing students per year in the baseline years are included. This is the sample applied in this study. The right-hand side figure illustrates the analytical sample created by the share criterion, where running variable is the number of low-performing students and only schools with a share of low-performing students above the regional cut-off are included.

Figure A2. RD first stage graph, Probability of participation in the RSA



APPENDIX B: Validation of regression discontinuity model

Figure B1. RD – density

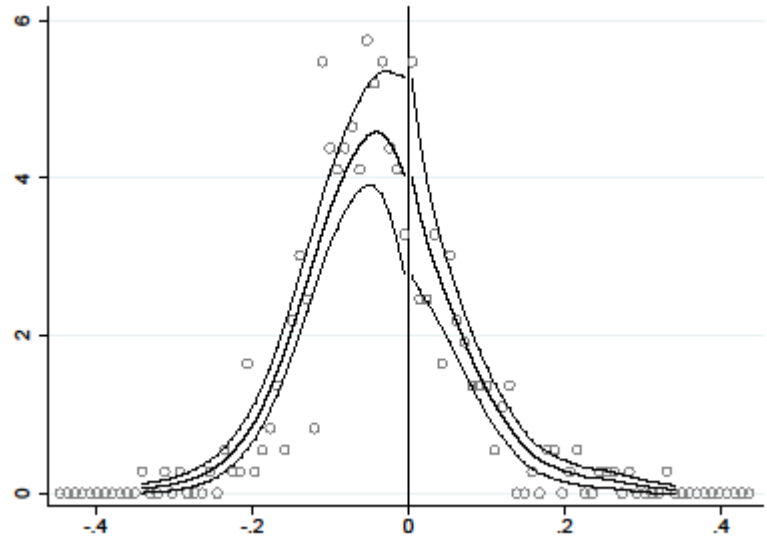


Table B1. Placebo regressions, RD results on pre-treatment outcome (2017)

RD Estimate	0.01	-0.00
	(0.02)	(0.02)
Polynomium	1	1
Bandwidth	.20	.20
Covar_Cutoff		X
n_eff	21154	21154
N	22503	22503

Clustered standard errors in parentheses
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Figure B2. RD balance graphs, predetermined covariates

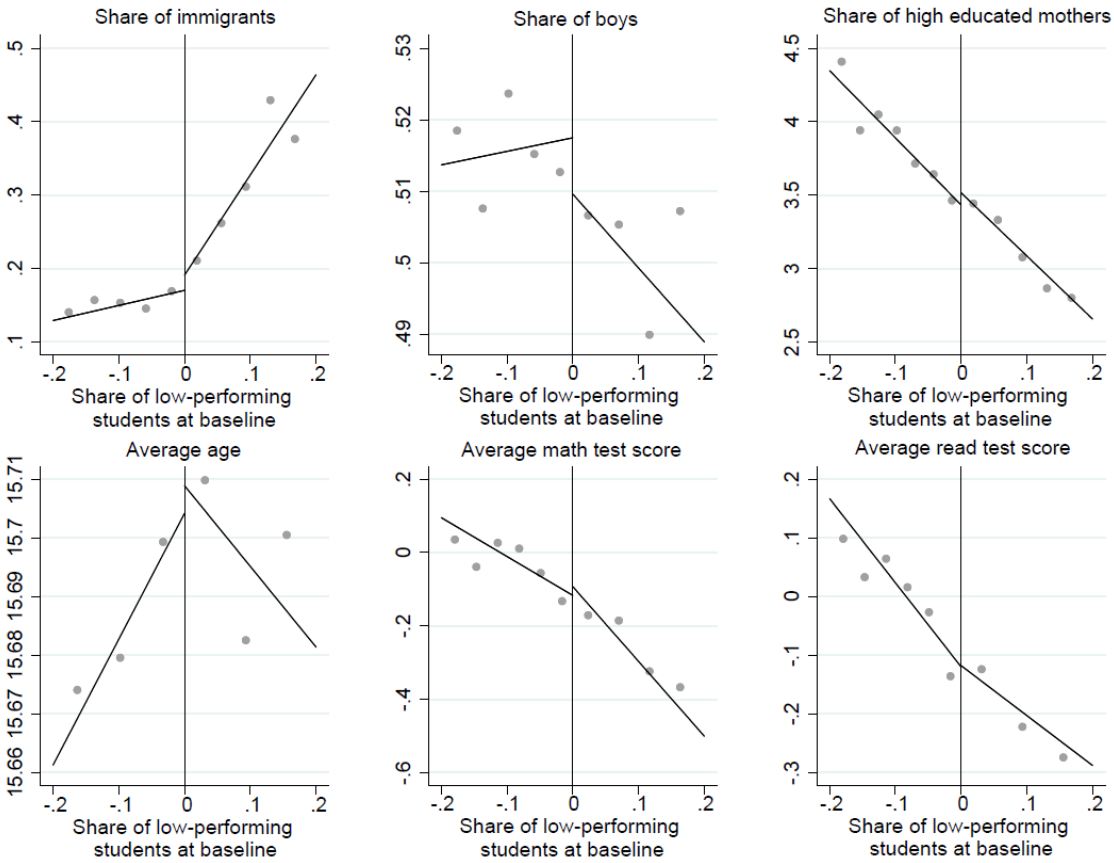


Table B2. RD balance regressions, predetermined covariates

	Immigrant	Boy	Age	Mathtest	Readtest	Mothers education
RD Estimate	0.02	-0.00	-0.01	0.04	0.04	0.03
	(0.03)	(0.01)	(0.02)	(0.05)	(0.05)	(0.02)
Polynomium	1	1	1	1	1	1
Bandwidth	0.20	0.20	0.20	0.20	0.20	0.20
Covar_cutoff						
n_eff	20506	20506	20506	19193	19206	19848
N	21888	21888	21888	20480	20493	21189

Clustered standard errors in parentheses
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

APPENDIX C: Unintended outcomes

Table C1. RD Results on the number of students in special classes

RD Estimate	-0.001	-0.01
	(0.02)	(0.01)
Polynomium	1	1
Bandwidth	.20	.20
Covar_Cutoff		X
n_eff	21533	21533
N	22959	22959

Clustered standard errors in parentheses
+ p < 0.10, * p < 0.05, ** p < 0.01

Table C2. RD Results on the probability of a student missing an exam

RD Estimate	-0.00	-0.00
	(0.01)	(0.01)
Polynomium	1	1
Bandwidth	.20	.20
Covar_Cutoff		X
n_eff	20506	20506
N	21888	21888

Clustered standard errors in parentheses
+ p < 0.10, * p < 0.05, ** p < 0.01

APPENDIX D: Robustness of difference in difference model

Table D1. DID matching

	Not matched	Matched
Control	525	172
Non-participating	0	18
Participating	12	91
Total	537	281

Table D2. DID balance table

	Immigrant	Boy	Age	Math test	Read test	Mothers education
DiD estimate	0.01	-0.02	-0.02	-0.01	-0.02	-0.01
	(0.011)	(0.02)	(0.02)	(0.04)	(0.04)	(0.01)
Observations	1385	1385	1385	1385	1385	1385

All models using matched sample. Clustered standard errors in parentheses. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table D3. DID - different number of strata

Number of strata	14	18	19	24	26
DiD estimate	-0.03**	-0.03**	-0.034***	-0.03**	-0.04**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Observations	2485	1615	1385	1660	1370

All models using matched sample. Clustered standard errors in parentheses. Covariatates and trend variable included but not shown.

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$