The Impact of Being Labeled as a Persistently Lowest Achieving School: Regression Discontinuity Evidence on School Sanctions

Guan Saw
Michigan State University

I-Chien Chen
Michigan State University

Barbara Schneider
Michigan State University

Ken Frank
Michigan State University

Venessa Keesler
Michigan Department of Education

Joseph Martineau
Michigan Department of Education

2013

The Impact of Being Labeled as a Persistently Lowest Achieving School: Regression Discontinuity Evidence on School Sanctions

Abstract

Since the No Child Left Behind Act was enacted, school sanction policies have been increasingly used as a means to incentivize failing schools to raise student achievement. Using state-wide high school data from Michigan, our regression discontinuity analyses show that the bottom 5% schools identified as Persistently Lowest Achieving (PLA), which are accompanied by threats of sanctions, increased their student achievement in reading and writing. No immediate impact on student performance in mathematics was detected after year one but its effects become noticeable in year two. We find no improvement in student achievement for those bottom 6-20% schools labeled as "watch list", which received no actual penalties, suggesting that labeling per se appears to have limited effects on school performance.

Keywords: persistently lowest achieving schools, school sanctions, labeling, regression discontinuity

The Impact of Being Labeled as a Persistently Lowest Achieving School: Regression Discontinuity Evidence on School Sanctions

Introduction

Since the implementation of the *No Child Left Behind Act of 2001* (NCLB) both federal and state governments have employed various sanctions for schools that fail to meet academic standards. The purpose of these sanctions is to incentivize schools to improve their students' academic achievement. Sanctioning policies tend to have a hierarchical structure whereby the initial phases are somewhat cautionary such as being placed on a "watch list," that become more intense if schools fail to improve within a set time frame. Some of these types of school sanctions include reconstitution, private management, conversion to a charter school, school closure, vouchers, and withholding funds (NCES, 2013a). By 2012, thirty-two states had imposed certain sanctioning policies on schools regardless of their Title 1 status (NCES, 2013b).

One could question whether school performance improves by employing a sanctioning plan that begins with actions that are relatively minor such as being put on a watch list with little resource implications before moving to more serious types of sanctions that have severe financial implications. It could be that "watch lists" and other types of labels create organizational change within the schools and/or raise attention and community pressure that can be effective in raising school performance level. Previous studies have investigated the effect of these types of reforms at the student level, assuming that sanction effects can be detected at the individual level (Chiang, 2009; Figlio & Rouse, 2006; Winters, Trivitt, & Greene, 2010). However, measuring outcomes at the student level can bias the effect of sanctions as low performing schools tend to have higher rates of student mobility and dropout (Lee & Burkam, 2003; Rumberger & Palardy, 2005). The very best students may leave when a school is labeled as lowest performing, on the

other hand, the lowest achieving students may be pushed out by the school (Haney, 2000; Heilig & Darling Hammond, 2008).

In this study, we use state-wide data from Michigan at the *school-level* to analyze the effects of being identified as a persistently lowest achieving (PLA) school in 2010 on school level student performance in two consecutive years. PLA schools are required to develop and implement a reform plan or face sanctions. Our analysis focuses on whether there is a significant difference in student performance over time between schools that have been placed on the PLA list in contrast to those schools with similar characteristics (including student achievement) not on the list. To conduct this analysis we employ a sharp regression discontinuity (RD) design which allows us to explore the discontinuity between being placed on a PLA list (i.e., the bottom 5%, or schools with a percentile ranking of less than 5) and a "watch list" (i.e., schools with a percentile ranking of five or higher, but lower than 20. Schools on the watch list were not subject to the same threat of sanctions at this time. By using statewide achievement-based school percentile rankings we are able to develop a more precise measure of assignment to the treatment (being placed on a list). As a falsification test, we use the state's criteria to construct a pseudo list of schools that would have been on "the list" in 2009 and estimate the effect on academic performance in 2010.

Evaluating the effects of the PLA in Michigan contributes to the relatively limited research on school sanctions. First, while a majority of previous studies focus on elementary and middle schools, this analysis is targeted at a statewide sample of high schools, which are generally considered to be more challenging in terms of turning around or restructuring. Second, this analysis evaluates the effects of being labeled as a PLA school for the first year of the policy as well as the subsequent year allowing for the examination of the policy's longer term effects.

Third, in 2009, the US Department of Education introduced the School Improvement Grant (SIG), a program for identifying and assisting the lowest 5% of schools; this study offers the first such evidence assessing the impact of the PLA list based on the criteria suggested by the SIG program. In the past, studies of school sanctions have tended to use Adequately Yearly Progress (AYP) status or grading policies implemented by the states. Using the PLA designation is a more comprehensive measure of school performance as it takes into account school performance in reading and mathematics and graduation rates for several years.

In recent years, many states have begun to identify and publicize the lowest achieving 5% schools and our empirical findings on the PLA list in Michigan suggest broad and timely applications for other school sanction systems. Our RD analyses indicate that the bottom 5% failing schools placed on the PLA list increased their student achievement scores in reading and writing after year one and these effects persisted in year two. While the positive effect in writing is quite robust, based on different estimation models, the positive effect in reading is less strong. Being on the PLA list has no immediate impact on student performance in mathematics for the first year but its effects become noticeable in year two. We find no impact on student achievement across academic subjects for those bottom 6-20% schools labeled as a "watch list school" and received no actual sanctions, suggesting that labeling per se appears to have no or limited effects on school performance.

Background

Labeling has been used for about twenty years as a means for incentivizing low performing schools to make substantive academic improvements. For example in Texas, schools were categorized as either "exemplary," "recognized," "acceptable" or "low-performing," or in Florida, as "A" through "F." These rating categories carried no immediate punitive actions as

they simply were labels. The NCLB Act in 2002 was a game changer, taking the labeling idea one step further by penalizing those failing schools. Beginning with NCLB, all states were required to conduct annual assessments and turn the results of those assessments into an AYP label of "made AYP" or "did not make AYP." Schools failing to make AYP were subject to a set of penalties, such as offering choice and transportation to another public school, but only after a school had failed to make AYP for multiple years. In the past few years, identifying the lowest achieving 5% of schools became a new school labeling practice across the nation, following the announcement of the SIG program in 2009 and the implementation of the Elementary and Secondary Education Act (ESEA) Flexibility in 2011, which gives the state flexibility from NCLB sanctions. Any states with an approved ESEA Flexibility application are required to identify the lowest 5% of schools (labeled Priority Schools) in their state (U.S. Department of Education, 2012).

The idea of identifying and publicizing low-performing schools is based in part on the notion that a negative label may stimulate a lowly-ranked school to make a change by stigmatized and shame-based motivation (Chakrabarti, 2013a; Chiang, 2009; Figlio & Rouse, 2006; Mintrop, 2004). It is often assumed that parents and teachers in schools identified as low performing would be more likely to move, especially if they do not have a reasonable voice in making improvements (Hirschman, 1970). While labeling may seem to be a reasonable strategy for triggering changes, the empirical findings have been mixed. Some researchers have found positive effects for school sanctions on the improvement of students' achievement after schools were being graded as failing schools. For example, in Florida, Figlio and Rouse (2006) and Rouse, Hannaway, Goldhaber, and Figlio (2013) find that schools receiving a grade of "F" improved their student performance immediately in the following year, particular in the high-

stakes tests in math and reading. They argue that schools actually changed in response to the increasing stigma and threat of vouchers. Other studies have documented similar positive effects of "F-rated" schools in Florida on student achievement in both high-stakes and low-stakes subjects, including math (Chiang, 2009; Winters, Trivitt, & Greene, 2010), writing (Chakrabarti, 2013a), reading and science (Winters, Trivitt, & Greene, 2010). Using data from New York City, which also adopted school letter grading and voucher policies similarly to those in Florida, Winters and Cowen (2012) find that schools that received "F" labels showed positive achievement improvement in reading and math, especially for those students in the bottom quartile.

While many studies suggest that school sanction policy led to increased academic achievement for students across varied subjects, a growing body of evidence reveals that some of this positive effect may be spurious. One of the reasons is that schools, under policy pressure in the form of sanctions, can boost the test scores by deliberately and systematically manipulating the population of students taking the high-stake tests (Chakrabarti, 2013b; Cullen & Reback, 2006; Figlio & Getzer, 2006; Haney, 2000; Heilig & Darling-Hammond, 2008). Heilig and Darling-Hammond (2008), for instance, find that in Texas schools strategically increased grade retention rates in the 9th grade and excluded more low-achieving students from taking the high-stakes state assessment in 10th grade. Figlio and Getzer (2006) also provided some evidence from Florida showing that schools tended to reclassify low-income and low-performing students as disabled or reassigned in special education categories that were exempt from the accountability system at the time.[1]

In short, despite a series of studies that have detected some positive relationship between

[1] Similar trends of increasing in the incidence of grade retention and reclassification of students as disabled were also documented in New York City (Allington & McGill-Franzen, 1992) and Chicago (Jacob, 2005) when a high-stakes testing accountability system was mandated.

school sanction policies and student performance, the conclusion is not without question. In addition, there are several limitations with these prior empirical studies which can be addressed by studying PLA designation in Michigan. First, the designations associated with the school labels or grades are oftentimes general. This has real implications when trying to assess the effect of a sanction on changes in performance. In Michigan, we can leverage the fact that schools on the PLA list (bottom 5%) or "watch list" (bottom 6-20%) are determined by cutoffs on a continuous measure of percentile ranking. Thus, we can employ a sharp regression discontinuity design with a precise measure of the assignment variable, which allows us to make a causal inference of the list effects. Second, in the above studies, researchers could not determine whether the effect of a label could be distinguished from a label which includes real sanctions such as allowing students to use vouchers and move to another school. In this study, we are able to compare two types of school labels, the watch list, in which there are no sanctions but only labels, with the PLA list, which makes the threat of sanctions a potential foreseeable outcome.

Persistently Lowest Achieving Schools in Michigan

Starting in 2010 the Michigan Department of Education (MDE), like many other states, published a state-wide school ranking list. Schools that fall in the bottom 5% of the percentile ranking, calculated by incorporating average achievement level and improvement rate both in math and reading, are classified as Persistently Lowest-Achieving (PLA) schools.[2] Schools with fewer than 30 students for each of the two most recent years with achievement data in both math

[2] MDE created the ranking used in this study in order to identify schools eligible for School Improvement Grants (SIG) and therefore had to follow specific ranking criteria for these grants developed by the U.S. Department of Education (USED). After the two initial years of using this ranking, MDE modified and enhanced the ranking and used that expanded ranking formula to identify PLA schools (now termed Priority Schools under Michigan's successful ESEA Flexibility application for the 2012-2013 school year). This study focuses on the impact of that original list and group of schools.

and reading are not eligible for a PLA designation. This study focuses only on the 346 regular high schools in the state that were eligible to receive a ranking in the 2010 list.[3] Of these 346 high schools, 19 were ranked below the 5% cutoff and identified as PLA schools. Once placed on the PLA list, schools were placed under the supervision of the School Reform officer and were required to develop and implement a reform plan that aims to rapidly increase student achievement. Those schools not making satisfactory progress are potentially subject to be taken over by the Statewide School Reform and Redesign District (SSRRD), meaning that the local community would lose governance of the community school identified as PLA (The Revised School Code Act, 2009).

Along with the PLA list (percentile rank less than 5) MDE also created a state watch list of schools in the next lowest 15% (percentile rank of five or higher and less than 20) which were identified as being in danger of becoming PLA schools in the future. In total, there were 57 high schools placed on the watch list in 2010. The watch list does not affect the PLA ranking; however, it provides an alert to the schools to avoid falling into the PLA category. There was no sanction imposed on these watch list schools, and no forecasted sanction for remaining on the watch list. Without a real threat of sanctions, those watch list schools may not be as responsive as their counterparts on the PLA list. Nonetheless, those schools being identified as watch list schools still received a label. While the label may only be symbolic, it still represents a "signal" of low performance and potential future sanctions in the case of declining school performance.

Methodological Approach

Sharp Regression Discontinuity Design

[3] All schools were potentially eligible to receive a ranking, but there was an overrepresentation in the original PLA list methodology used in 2010 and 2011 on a) high schools and b) Title I schools. These requirements have been altered in the methodology Michigan now uses to identify Priority (lowest 5%) schools and now all schools are equally eligible.

To identify the causal impact of being on the PLA or watch list in Michigan we use regression discontinuity designs. We estimate unbiased effects of being on the low-performing list by examining the difference in student achievement at the school level between schools just below and above a fixed threshold, (the fifth percentile for the PLA list, and the 20th percentile for the watch list). The assignment to a list is determined by the value of the percentile ranking on either side of a single cutoff point where there is perfect compliance of treatment. Thus we use sharp RD designs (Hahn, Todd, & van der Klaauw, 2001; Trochim, 1984) and follow the procedures laid out by Imbens and Lemiuex (2008).

First, we illustrate the regression discontinuity identification strategy by running local linear regressions on both sides of the fixed cutoff and plot the relationship between school outcomes and percentile ranking. The major purpose of the nonparametric graphical analysis is to explore whether there is a jump in the conditional mean of the outcome around the cutoff as an evidence of a discontinuity. For our PLA list study, which only involves a small size of school samples, we use all 19 PLA schools below the cutoff (bottom 5%) and 19 non-PLA schools above the cutoff (bottom 6%-10%) in the local linear regressions. This selection of our bandwidth is critical for implementing local methods when analyzing such a small finite sample as suggested by Lee and Lemiuex (2010).

Second, we formally evaluate the effect of being on a list by estimating parametric regression discontinuity models. To carry out the estimation, we let Y_{i1} be the school level student achievement in a given subject for a school *i* in year 2011, $List_{i0}$ be a dummy variable for whether school *i* was placed on the PLA or watch list in year 2010 (as the probability of being placed on a list jumps from 0 to 1 at the cut score), and $PctRank_{i0}$ be the percentile rankings rated by school *i* in year 2010. The RD estimation model is represented by the

following equation:

$$Y_{i1} = \beta_0 + \beta_1 List_{i0} + \beta_2 PctRank_{i0} + \varepsilon_{i1} \quad (1)$$

where the estimated coefficient β_1 is the effect of being on the list. On one hand, we expect to see positive PLA list effects as a result of the intensified labeling process and the real threat of sanctions. On the other hand, we expect the effect to be limited or non-existent for being on the watch list as those schools were only going through relatively moderate labeling processes without any real threat of sanctions.

One of the most critical aspects of the RD modeling is the functional form specification of the relationship between the forcing and outcome variables (Schochet et al., 2010). Using an incorrect functional form in RD designs typically produces a bias in treatment effect (Lee & Lemieux, 2010). In our case, we assume the association between percentile ranking and school outcomes was linear given the percentile ranking is also computed based on school performance in previous years. Yet, the functional form can be nonlinear due to a nonlinear relationship between the two variables and interactions which may occur between the forcing variable and treatment. Thus, we test a variety of functional forms by including quadratic, cubic, and interaction terms into equation (1) to determine which best fits the data. These specifications did not lead to improvements in model fit so we only present estimation results from the more parsimonious linear specification.

We further address the limitation of small sample size in our study by adding a vector of selected school characteristics surveyed in year 2010, X_{i0}, to the regression function (1) to eliminate any small sample biases and to improve the precision of our estimation (Imbens & Lemieux, 2008). These school level factors include percentage of free/reduced lunch students,

percent of minority students, school size, and pupil teacher ratio.[4] Together with the percentile ranking as the forcing variable, these covariates are assumed and also have been empirically tested not to be influenced by the treatment (see the results of unconfoundedness assumption tested in Appendix A). The extended RD estimation model with covariates is represented by the following equation:

$$Y_{i1} = \beta_0 + \beta_1 List_{i0} + \beta_2 PctRank_{i0} + \gamma X_{i0} + \varepsilon_{i1} \quad (2)$$

In sharp RD designs, inferring a causal impact on an outcome relies on some fundamental assumptions including: (1) the jump at the cut score is truly discontinuous; (2) the forcing variable is observed without error; (3) the dependent measure, in this case the achievement score, is a continuous function of the assignment variable (percentile ranking) at the cutoff in the absence of treatment, and (4) the treatment units are sorted unconditionally by assignment. We conduct a series of tests regarding violations of the above assumptions, recommended by Imbens & Lemieux (2008), including unconfoundedness, no-manipulation, and no jumps at non-discontinuity points. Results from testing these identification assumptions are presented in Appendix A and generally support the use of the sharp regression discontinuity design with our data and specifications described in equations (1) and (2).

To assess the robustness of our primary RD estimation results, we compare the 2010 PLA list effects to estimated results of a pseudo 2009 PLA list which is constructed by using data from the previous school years before the state mandated assignment for low-performing schools to a PLA list. Similar falsification tests or placebo tests are often used in policy analysis, especially when evaluating intervention programs which reward or penalize schools based on

[4] The school level covariates are obtained from the Common Core Data (CCD) provided by U.S. Department of Education's National Center for Education Statistics (NCES). They are collected in the 2009-2010 academic year. Appendix Table A1 reports descriptive statistics of the four variables for PLA and watch list samples.

students' average performance. By applying the same estimation models to a historical counterfactual school sample before the PLA list was implemented, we expected to verify that there would be no effect of "the list" on achievement outcomes, since none of the schools in the this pseudo treatment group (bottom 5% in 2009) would have experienced labeling and threat of sanctions from being on a "persistently lowest achieving" school list.

Data and Measures

We use school-level data provided by the Michigan Department of Education to examine the effect of the 2010 PLA list on school outcomes measured in 2011 and 2012. The school data contain all criteria used for determining the 2010 PLA and watch list, including information on percent proficient in math and reading in statewide high school examinations (Michigan Merit Examination, MME) for the past four years (from 2006 to 2009), number of students tested in math and reading for the past four years, calculated four year improvement slopes in math and reading, graduation rates for the most recent three years, percentile ranks, and whether the school was placed on the PLA or watch list. The percentile ranking, which is the forcing variable in our RD estimation models, is a continuous rating score that is positively correlated with school outcomes. This linear relationship is assumed and has been empirically tested to be smooth, hence any discontinuity of the conditional distribution of the school outcomes as a function of the percentile ranking at the cut score is considered as evidence of a causal effect of being on the PLA list.

The outcome variables are drawn from student achievement on the Michigan Merit Exit (MME) assessed at the end of the 2010-2011 and 2011-2012 school years.[5] Based on Michigan

[5] One limitation of our study is that the school level outcome data are in some ways different from the data with student level information that MDE used for accountability purposes. One key aspect is that MDE used only achievement scores of students who have been at a school for at least one Full Academic Year (FAY) in the percentile ranking computation. Our data include all tested students regardless of their FAY status. An internal

high school standards, the MME is administered annually in the spring to high school juniors. Five subjects are tested, including mathematics, reading, writing, science, and social studies. Student performance falls into one of four categories: advanced, proficient, partially proficient, and not proficient. Students who score either "proficient" or "advanced" are considered as having met the proficiency level in a specific subject, which is a critical component used in computing state-wide school percentile rankings. We use percent of students who exceeded proficiency levels in the reading, writing, and mathematics for each school of 2011 and 2012 as the primary school outcome variables.[6]

Table 1 reports the mean percent of students who met proficiency levels in the three subjects from 2009 to 2012 for the 2010 PLA list and watch list separately. Of the three subjects, over the years, reading has the highest percentage of students who were at least proficient, writing the second highest, whereas mathematics the lowest. For the PLA list schools, for instance, the percentage of students who were at least proficient in reading in this time frame ranged from 27.2% to 32.9%, in writing from 20.5% to 26.4%, and in math from 5.9% to 8.5%. Further, in every group of PLA and watch list schools, while the average proportion of students who met proficiency levels in reading and writing increased consistently from 2009 to 2012, in math the numbers dropped slightly in 2010 before increasing in the following years.

Insert Table 1 here

Results

Effects of Being on PLA List

evaluation carried out by MDE found that there is no substantial difference in school performance regardless of whether FAY status is taken into account.

[6] In our additional analyses, we find that the effects of the PLA list on science and social studies show a similar pattern as those on reading, writing, and mathematics (results available upon request).

In Figure 1, we illustrate the relationship between the assignment variable of percentile ranking and school-level performance in three different subjects (e.g., reading, writing, and mathematics) in 2011.[7] Any observed upward jump at the cutoff can be interpreted as the positive effect of PLA list on school achievement. To quantify the magnitude and significance of the discontinuities in school outcomes due to being on the list, we run local linear regressions or "kernel" regression models on both sides of the fixed threshold. Given such a small sample size in our study, we chose to include all 19 PLA list schools and a comparable number of 19 non-PLA (watch list) schools in the estimations. As expected, since we use all the school samples, results from the local linear regressions approximate those estimated by running parametric regression discontinuity models.

Insert Figure 1 here

The formal estimation results of parametric RD models based on specification (1) and (2), which use subject-specific outcomes, are reported in Table 2. The first column for every subject in Table 2 displays the estimated first stage impact of being on the 2010 PLA list for the percentage of students who met proficiency levels in 2011. Both unconditional and conditional (with covariates) estimates clearly indicate positive statistically significant effects of the PLA list on reading and writing at the critical level of 5% (based on one-tailed tests)[8]. The unconditional model estimates show that being placed on the PLA list boosts the percentage of students who met the proficiency level in reading by 13.0 percentage points (p=0.026) and in writing by 18.3

[7] The figures showing the relationship between percentile ranking and school outcomes in the three subjects in 2012, which are not presented here, look very similar to Figure 1. However, they have slightly different size of upward gap at the cutoff.

[8] Given our research hypothesis which is being placed on PLA list positively increase student performance is directional and such a small sample size, we assess the statistical significance of estimates based on one-tailed tests. For watch list and pseudo-list analyses, two-tailed tests are employed as we do not have a directional hypothesis on the possible effects.

percentage points (p=0.002) in 2011. Furthermore, all estimates on the running variable and covariates in every model are in the expected direction (coefficients available upon request).

Insert Table 2 here

In Table 2, the second column reports the causal estimates of the PLA list effect on school performance for the percent of students who met proficiency levels in the three subjects in 2012. While the magnitude and significance are generally weaker in reading, and writing, we find that the sustained effect of the PLA list occurs across subjects. Based on the unconditional models, being on the 2010 PLA list significantly increases the proportion of students who met proficiency levels in reading by 11.9 percentage points (p=0.049) and in writing by 13.5 percentage points (p=0.035) in 2012. Furthermore, the size of the estimated coefficient in math slightly increased from 2011 to 2012 and achieves statistical significance at the critical level of 5% in the conditional model.

One crucial concern about studying effects of school sanctions is that schools might strategically manipulate the population of test-takers, particularly grade retention and student exclusion (Figlio & Getzer, 2006; Heilig & Darling-Hammond, 2008). We address this issue by examining the student attrition rates for 2010 tenth graders that were promoted to eleventh grade in 2011 and proportion of 11^{th} graders in a school taking the MME test in 2011, in both PLA and non-PLA schools. Our data show that on average PLA schools lost a relatively higher percentage of students (15.9%) for the studied student cohort, compared with non-PLA schools (10.5%), yet the difference is not statistically significant. Additionally, we find no evidence that there is a significant higher percentage of 11^{th} graders in PLA schools (12.7%) being held from taking the high-stake tests, as compared with non-PLA schools (10.8%).

Another concern that has been raised when evaluating outcomes of accountability policies is that schools under pressure may only focus their efforts and resources on improving

certain student subgroups that are just below the cutoff (Chakrabarti, 2013a, 2013c; Ladd & Lauen 2010; Neal & Schanzenbach, 2010). In this study, we address the issue by calculating percent of students who were at least partially proficient (who were graded partially proficient, proficient, or advanced) in the three subjects at the school level in 2011. Appendix Table A3 presents the results, showing that for lowest performing schools enrolled with a majority of low-achieving students, being on the PLA list boosts the percentage of students who were at least partially proficient in the following year. These results provide additional evidence supporting the positive effects of being on the PLA list for both students near and far below proficiency.

Effects of Being on Watch List

The above RD estimation results using first year outcomes of student achievement suggests that those schools placed on the PLA list improve student performance even before a formal implementation of a reform plan took place. However, the precise interpretation of the findings proves challenging as we cannot distinguish whether the positive effects are due to the practice of stigmatized labeling or threat of sanctions faced by these schools. In an attempt to shed light on the relative importance of labeling and sanction threat effects, we estimate the causal impact of being on the 2010 watch list. Without any real threat of sanctions, those schools that were being placed on the watch list are only going through a labeling process.

Table 3 presents the results from estimating equation (1) using the 2010 watch list school samples. Only the estimated coefficients on school-level performance of percent of students who met proficiency levels in 2011 using the full samples of watch list schools (a bandwidth of 15% above and below the fixed threshold) are reported. We find no evidence suggesting that there is a significant gap in achievement between watch list schools and those comparable school samples. Magnitudes of all estimates are modest (or much smaller than those PLA list estimates in Table

2) and none of them reach the level of statistical significance. Moreover, unlike the estimation of the PLA list effect which yields all estimates in positive direction, no clear pattern exists across subjects in watch list effect estimations. Estimations using subsamples with different bandwidths of 10% and 5% do not yield any notable changes in size and significance level for all estimates. These additional estimation results are presented in Table B3 in Appendix B.

Insert Table 3 here

Effects of Being Pseudo-PLA List

Our primary results show that being on the 2010 PLA list increases the percentage of students who met proficiency levels in several subjects. This finding raises the question whether the estimated positive effects are due to the tendency of the lowest-performing schools in a specific year to revert back to its normal performance level in student achievement the following year. As a robustness check, Table 4 presents the results from estimating the effects of being placed on a pseudo-2009 PLA list which is created using data prior to 2010. We find no significant effect of the pseudo-2009 PLA list on any academic subjects, suggesting that the bottom 5% schools in 2009, which would have been on a PLA list if the policy was being enforced one year earlier, did not substantially improve student performance in the following year. Unlike the results from our 2010 PLA list analyses, which show all estimates are in positive direction and many of them have sizable coefficients, magnitudes for all estimates from the pseudo-2009 PLA list estimations are negligible and the directions are mixed.

Insert Table 4 here

In addition to the above falsification test of the 2009-pseudo PLA list analysis, we employ two additional robustness check strategies. First, as the selection of bandwidth of observations for estimation is one of the important statistical issues in RD designs (Imbens & Lemiuex, 2008; Schochet et al., 2010), we investigate the sensitivity of our primary RD

estimation results by conducting analyses using differences in bandwidths. In Appendix Table A5, we find that our estimation of the positive effect of the PLA list on writing is quite robust to other bandwidth selections, however, it is less so for reading. Second, we compare school achievement in the following years between the PLA list schools and those schools that would have showed up on the PLA list if the student sample size restriction, which is having at least 30 full academic year students with mathematics and reading scores in the most recent two years, were removed. These are unique school samples that can serve as control units for examining the treatment effect of the PLA list.[9] Appendix Table A6 shows descriptive statistics of these small schools and Table A7 displays the results, showing that the PLA list schools perform better in improving student performance across subjects than those small schools do. Taken together, we find no reasons to doubt the robustness of our primary findings of positive causal effects of being placed on the 2010 PLA list in Michigan.

Discussion

This study provides new evidence on school sanctions by assessing the effects of being on the PLA list among traditional high schools in Michigan, To summarize, our RD estimations suggest that being on the 2010 PLA list increases the percent of students who met proficiency levels in reading and writing, by ranging roughly from 12 to 18 percentage points, in the two subsequent years. The positive effects in writing are quite robust but are less in reading, based on estimating additional models using control variables, a new set of outcomes, different bandwidths, and alternative analysis samples. In addition, we find that the PLA list status has no impact on math achievement immediately after one year but its magnitudes of effects are noticeable in the following year.

[9] These small schools are identical to the PLA list schools in terms of student performance in standardized tests and graduation rates, which are two major criteria in calculating state-wide achievement base percentile rankings.

Taking advantage of the unique context in Michigan that there are two types of low-performing school lists which were constituted in the same year, we explored how schools would respond to different levels of labeling processes and accountability pressures. Unlike meaningful positive effects demonstrated by the PLA list schools, our analyses show that there is no remarkable difference between watch list schools and comparable schools in terms of student performance in every examined subject, including reading, writing, and math, one year later. Similar conclusions were reached when estimating the models by including control variables or using different choices of bandwidths.

Previous studies have shown that schools, when facing increased accountability pressure, may strategically reallocate instructional resources and teachers' time on subjects which are easier to improve in the short term (Smith, Roderick & Degener, 2005). Goldhaber and Hannaway (2004), and Chakrabarti (2013a), for example, find that students in failing schools are more likely to have the biggest score gain in writing, which is considered as one of the subjects that students can improve quickly. We also find that being on the PLA list tends to have a strong positive impact on student achievement in writing but modest effects in other subjects, including reading and social studies (results not reported here). We recognize that the observed immediate strong positive effects in writing may be a result of test preparation coaching at the early stage of sanction (Darling-Hammond & Wise, 1985). Furthermore, student performance in writing is included in the calculation of percentile ranking in the following years. PLA schools, therefore, may become aware that they need to focus on improving writing scores.

Furthermore, some argue that schools under accountability pressure tend to produce immediate increases in test scores, but often ignore possible reform strategies that can lead to longer-term improvement (Winters & Cowen, 2012; Chiang, 2009). Utilizing several years of

data in Michigan, we find some evidence indicating that for student achievement in math, there is no significant gain after one year of being placed on the PLA list, but it appears to improve in the subsequent year. Our findings are consistent with Winters and Cowen's (2012) study drawing data from New York City, which reports that the gain in student math scores was observed two years after schools received an "F" grade. Unlike writing, it may take a longer period and more teaching resources to achieve the gain of student achievement in mathematics.

Previous research seeks to understand whether it is stigma or threat of sanction behind the school sanctioning process that serves as a major driving force to make school change (Figlio & Rouse, 2006; Chakrabarti, 2013a). While our data cannot distinguish the two factors, we can shed some light on the issue by analyzing and comparing the effects of the two types of low-performing school lists (i.e. PLA and watch list) simultaneously. We fail to find systematic evidence that schools being on the watch list improve student achievement across academic subjects. This suggests that stigmatization or labeling per se appears to not be a possible triggering factor for failing schools to make change. It is important to note that as the education authorities and local media pay less attention to these watch list schools, the intensity of the labeling process is relatively low.

The overall findings of our study are notable for the various reasons discussed above, yet we recognize that there are a number of caveats and nuances. First, there are several limitations of this study both in terms of analytical approach as well as data set. The control group is selected from the cluster of schools that ranked between 5% and 10% on the percentile list, and the treatment schools ranked all lower than 5%. The assumption here was that the two populations are similar and comparable at a certain degree. However, this assumption could be violated considering the difficulty of improving school performances consistently at the bottom.

Particularly, when the bottom schools are characterized as enrolling consistently disadvantaged students in terms of educational resources, it might be difficult to find differences between the treatment and control schools as they both serve similar populations with similar resources. To make the two populations comparable, we needed to narrow the window which might cause attenuated regression coefficients due to the limited range of variance. On the other hand, too wide a window can also be problematic in that it can generate non-comparable population groups. Therefore, a causal inference from the regression discontinuity design inevitably relies on the belief that we made a comparable control group using the range of a narrow window, and this belief may be questionable. In that sense, this study cannot be fully free from the limitation of a regression discontinuity design.

The limited variables in the data set also leave the following concerns. First, we lack implementation information. We do not have the proper information to investigate what actual changes are being implemented in the schools in the first and second year after the PLA list announcement. Second, the school level variables are quite limited in content and there may be other factors such as change in school leadership or intensive professional development that could be affecting the performance in the schools. To achieve a more definitive conclusion, such variables should be considered.

Nonetheless, even with these few variables and a short-time, comparing a narrow band of similar schools, PLA list status appears to have some positive effects. It seems prudent to investigate these types of sanctions as they are relatively cost efficient. Moreover, understanding how schools respond to specific rules of accountability is crucial for designing effective school reform programs. To date, following the implementation of the SIG program and ESEA Flexibility, many states have started to identify and publish the lowest performing 5% school list

annually. Thus our findings from studying the PLA and watch list in Michigan may have a broad application to the similar systems or school sanction policies that are becoming prevalent across the nation. In this paper, we provide evidence suggesting that once being identified and placed on the PLA list, which is accompanied by a foreseeable threat of sanctions, schools may respond positively to improve student performance. However, simply putting a warning label or stigma on a failing school appears to not be sufficient to create a positive response toward change.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research has been supported by the Michigan Consortium for Educational Research (MCER) through Institute of Education Sciences, U.S. Department of Education, under Grant No. R305E1-00008. The opinions expressed are those of the authors and do not represent views of the Institute of Education Sciences or the U.S. Department of Education.

References

Allington, R., & McGill-Franzen, A. (1992). Unintended effects of educational reform in New York State. *Educational Policy, 6,* 396-413.

Chakrabarti, R. (2013a). Vouchers, public school response and the role of incentives: Evidence from Florida. *Economic Inquiry, 51*(1), 500-526.

Chakrabarti, R. (2013b). Accountability with voucher threats, responses and the test-taking population: Regression discontinuity evidence from Florida. *Education Finance and Policy 8*(2), 121-267.

Chakrabarti, R. (2013c). Incentives and responses under No Child Left Behind: Credible threats and the role of competition. *Journal of Public Economics.* http://dx.doi.org/10.1016/j.jpubeco.2013.08.005

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics, 93*(9-10), 1045-1057.

Cullen, J., & Reback, R (2006). Tinkering towards accolades: School gaming under a performance accountability system. In T. J. Gronberg & D. W. Jansen (Eds.), *Improving school accountability: Check-ups or choice (Advances in Applied Microeconomics, Volume 14)* (pp. 1-35). Amsterdam: Elsevier Science.

Darling-Hammond, L., & Wise, L. (1985). Beyond standardization: State standards and school improvement. *The Elementary School Journal, 85*, 315–336.

Figlio, D. N., & Getzler, L. (2006). Accountability, ability, and disability: Gaming the system? In T. J. Gronberg & D. W. Jansen (Eds.), *Improving school accountability (Advances in Applied Microeconomics, Volume 14)* (pp. 35-49). Amsterdam: Elsevier Science.

Figlio, D. N., & Rouse, C.E. (2006). Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics, 90,* 239– 255.

Goldhaber, D., & Hannaway, J. (2004). Accountability with a kicker: Observations on the Florida A+ Accountability Plan. *Phi Delta Kappan, 85*(8), 598-605.

Hahn, J., Todd, P., & van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica, 69*, 201–209.

Haney, W. (2000). The myth of the Texas miracle in education. *Education Policy Analysis Archives, 8* (41). Retrieved from: http://epaa.asu.edu/ojs/article/view/432/828

Heilig, J. V., & Darling-Hammond, L. (2008). Accountability Texas-style: The progress and learning of urban minority students in a high-stakes testing context. *Educational Evaluation and Policy Analysis, 30*(2),75-110.

Hirschman, A. O. (1970). *Exit, voice, and loyalty: Responses to decline in firms, organizations, and states*. Cambridge, Mass.: Harvard University Press.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics, 142*(2), 615-35.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics, 89*(5-6), 761-796.

Ladd, H. F., & Lauen, D. L. (2010). Status versus growth: The distributional effects of school accountability policies. *Journal of Policy Analysis and Management, 29*(3), 426-450.

Lee, V. E., & Burkam, D. T. (2003). Dropping out of high school: The role of school organization and structure. *American Educational Research Journal, 40*, 353–393.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature, 48*, 281–355.

Mintrop, H. (2004). *Schools on probation: How accountability works (and doesn't work).* New York: Teachers College.

National Center for Education Statistics (2013a). *State education reforms: Table 1.4. Types of school sanctions, by state (2011-2012).* Retrieved from: http://nces.ed.gov/programs/statereform/tab1_4.asp

National Center for Education Statistics (2013b). *State education reforms: Table 1.3. Rewards and sanctions for schools, by state (2011-2012).* Retrieved from: http://nces.ed.gov/programs/statereform/tab1_3.asp

Neal, D., & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *Review of Economics and Statistics, 92*(2), 263-283.

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.

Rumberger, R. W., & Palardy, G. J. (2005). Test scores, dropout rates, and transfer rates as alternative indicators of high school performance. *American Educational Research Journal, 42*(1), 3-42.

Schochet, P., Cook, T., Deke, J., Imbens, G., Lockwood, J. R., Porter, J., & Smith, J. (2010). *Standards for regression discontinuity designs*. Retrieved from : http://ies.ed.gov/ncee/wwc/pdf/wwc_rd.pdf.

Smith, B. A., Roderick, M., & Degener, S. C. (2005). Extended learning time and student accountability: Assessing outcomes and options for elementary and middle grades. *Educational Administration Quarterly, 41*(2), 195-236.

The Revised School Code Act, 451 of 1976, §§ 380-1280c, Laws of Michigan, 2009.

Trochim, W. M. K. (1984). *Research design for program evaluation: The regression discontinuity approach*. Beverly Hills: Sage.

U.S. Department of Education. (2012). *ESEA flexibility*. Washington, DC: Author. Retrieved from http://www.ed.gov/esea/flexibility

Winters, M. A., Trivitt, J. R., & Greene, J. P. (2010). The impact of high-stakes testing on student proficiency in low-stakes subjects: Evidence from Florida's elementary science exam. *Economics of Education Review, 29*(1), 138-46.

Winters, M. A., & Cowen, J. M. (2012). Grading New York: Accountability and student proficiency in America's largest school district. *Educational Evaluation and Policy Analysis, 34*(3), 313-327.

TABLE 1
Mean Percent of Students who met Proficiency Levels in Reading, Writing, and Mathematics 2009-2012 by Percentile Ranking

	% of students met proficiency level							
	2009		2010		2011		2012	
	Percentile rank		Percentile rank		Percentile rank		Percentile rank	
PLA list sample	<5% (*n*=19)	6-10% (*n*=19)	<5% (*n*=19)	6-10% (*n*=19)	<5% (*n*=19)	6-10% (*n*=19)	<5% (*n*=19)	6-10% (*n*=19)
Reading	27.21 (11.70)	36.55 (8.25)	29.31 (12.10)	37.16 (8.91)	29.42 (12.80)	38.19 (10.37)	32.88 (13.88)	43.39 (10.76)
Writing	21.24 (10.98)	27.62 (7.74)	20.46 (9.68)	26.46 (9.09)	24.65 (13.26)	29.86 (9.75)	26.38 (12.55)	35.61 (13.06)
Mathematics	8.10 (5.83)	12.23 (4.93)	5.87 (4.98)	11.63 (6.31)	6.73 (5.59)	11.39 (5.92)	8.53 (6.71)	13.41 (6.64)
Watch list sample	6-20% (*n*=57)	21-35% (*n*=56)	6-20% (*n*=57)	21-35% (*n*=56)	6-20% (*n*=57)	21-35% (*n*=56)	6-20% (*n*=57)	21-35% (*n*=56)
Reading	40.13 (8.37)	45.10 (7.74)	42.41 (8.47)	49.84 (6.32)	43.65 (11.62)	51.47 (8.64)	47.29 (10.90)	53.15 (9.84)
Writing	32.28 (8.91)	37.25 (8.96)	31.42 (9.19)	37.41 (7.48)	35.15 (10.76)	42.96 (9.16)	39.46 (12.31)	44.52 (11.23)
Mathematics	14.27 (6.25)	19.90 (7.47)	13.47 (5.64)	18.56 (8.13)	14.92 (7.43)	21.13 (8.11)	16.36 (7.79)	21.87 (9.17)

Source: Michigan Merit Examination (MME), Michigan Department of Education (MDE).
Note: *n* = sample size. High school mean achievement is reported in percent of students met proficiency level in a given subject with the standard deviation in parentheses.

TABLE 2
Estimated Causal Effects of Being on the 2010 PLA List (h=5%, n=38)

	% of students met proficiency level					
	Reading		Writing		Mathematics	
	2011	2012	2011	2012	2011	2012
Unconditional model	13.040 *	11.910 *	18.338 **	13.506 *	2.973	5.048
	(6.454)	(6.985)	(6.207)	(7.250)	(3.515)	(3.983)
Conditional model	10.064 *	9.333 †	15.854 **	10.847 †	2.510	6.467 *
(with covariates)	(5.856)	(5.655)	(6.171)	(6.555)	(3.748)	(3.635)

Note. h = bandwidth; n = sample size. Taken from a separate regression model on schools, each cell in the table shows the estimated coefficient on a dummy variable indicating the effect of being on the 2010 PLA list. All estimation models include the forcing variable of percentile ranking as a predictor. The covariate variables, including percent of free/reduced lunch students, percent of minority students, school size, and pupil teacher ratio, are collected in the 2009-2010 academic year. Standard errors are in parentheses. Statistical significance is determined using one-tailed tests.
*** p<.001; ** p<.01; * p<.05; †p < .10.

TABLE 3
Estimated Causal Effects of Being on the 2010 Watch List (*h*=15%, *n*=113)

	% of students met proficiency level					
	Reading		Writing		Mathematics	
	2011	2012	2011	2012	2011	2012
Unconditional model	0.779	0.662	-0.399	3.267	-1.832	0.269
	(3.703)	(3.807)	(3.637)	(4.300)	(2.859)	(3.106)

Note. h = bandwidth; n = sample size. Taken from a separate regression model on schools, each cell in the table shows the estimated coefficient on a dummy variable indicating the effect of being on the 2010 Watch list. All estimation models include the forcing variable of percentile ranking as a predictor. Standard errors are in parentheses. Statistical significance is determined using two-tailed tests. The conditional models which include covariates of percent of free/reduced lunch students, percent of minority students, school size, and pupil teacher ratio, shows similar results in this table.
*** p<.001; ** p<.01; * p<.05; †p < .10.

TABLE 4
Robustness Check: Effects of Being on the 2009 pseudo-PLA List (h=5%, n=37)

	% of students met proficiency level in 2010		
	Reading	Writing	Mathematics
Unconditional model	0.688	-3.600	4.107
	(4.499)	(4.699)	(3.007)

Note. h = bandwidth; n = sample size. Taken from a separate regression model on schools, each cell in the table shows the estimated coefficient on a dummy variable indicating the effect of being on the 2009 pseudo-PLA list. All estimation models include the forcing variable of percentile ranking as a predictor. Standard errors are in parentheses. Statistical significance is determined using two-tailed tests. The conditional models which include covariates of percent of free/reduced lunch students, percent of minority students, school size, and pupil teacher ratio, shows similar results in this table.
*** p<.001; ** p<.01; * p<.05; †p < .10.

Figure Captions

Figure 1. *Nonparametric graphs of the relationship between percentile rank in 2010 and school outcomes in reading, writing, mathematics in 2011.*

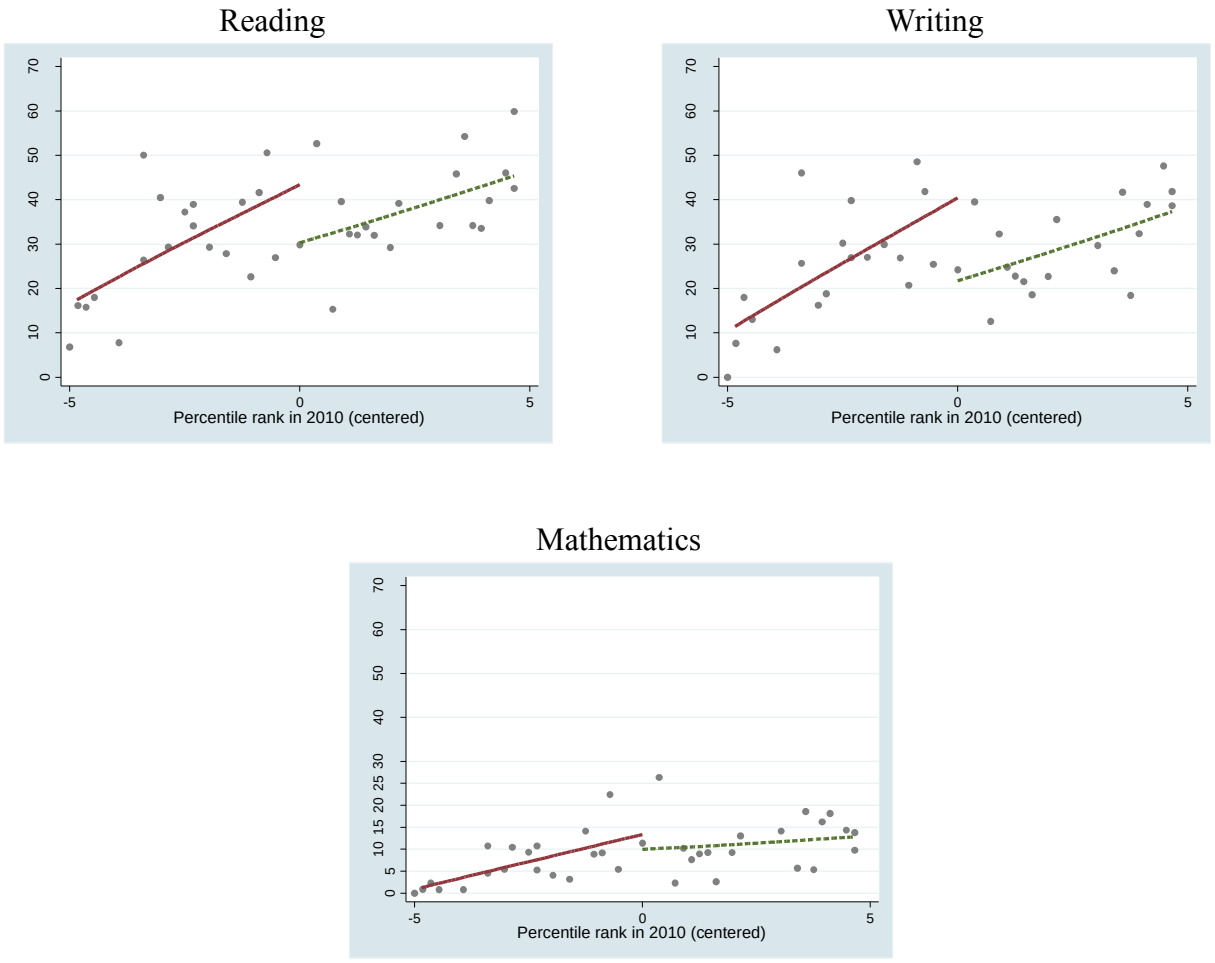

FIGURE 1. *Nonparametric graphs of the relationship between percentile rank in 2010 and school outcomes in reading, writing, mathematics in 2011.*

Appendix: Supplementary Data and Analyses (can be made available only on-line)

1. Descriptive Statistics of School Covariates

TABLE A1
Descriptive Statistics of School Covariates in 2010 by Percentile-Rankings

	Percentile ranking in 2010			
	PLA list study		Watch list study	
	<5% (*n*=19)	6-10% (*n*=19)	6-20% (*n*=57)	21-35% (*n*=56)
	mean (sd)	*mean (sd)*	*mean (sd)*	*mean (sd)*
% of free/reduced lunch students	63.67 (14.53)	55.26 (12.18)	50.80 (14.20)	45.14 (12.89)
% of minority students	60.27 (35.42)	35.40 (29.63)	28.59 (27.97)	12.30 (13.17)
School size	619.68 (283.02)	778.74 (421.29)	762.21 (483.63)	647.25 (487.45)
Pupil teacher ratio	17.70 (2.85)	18.70 (2.84)	19.21 (2.93)	19.21 (2.43)

Source: Common Core Data (CCD), National Center for Education Statistics, U.S. Department of Education.

2. Testing of Sharp RD Assumptions

(a) *Unconfoundedness assumption*. First, we examine the unconfoundedness assumptions by testing the null hypothesis of a zero average effect on other school characteristics as pseudo outcomes known not to be impacted by the treatment (Lee et al., 2004). We expect to find nojumps in the value of school measures at the cutoff point, which may invalidate the regression discontinuity design. As suggested by Lee and Lemieux (2010), it is useful to perform a seemingly unrelated regression (SUR) analysis if there are multiple covariates available. Our SUR model consists of the RD basic specification in equation (1) but with a different school covariate served as the outcome. Then, we perform a chi-square test for testing the hypothesis that all discontinuity terms across the four covariates are jointly equal to zero. Results yield a chi-square value of 7.99, with 4 degrees of freedom, and a p-value of 0.0921. Therefore we cannot reject the null hypothesis of no discontinuities of covariates around the fixed threshold.

(b) *No-manipulation assumption*. Second, we assess the no-manipulation assumption which suggests that no individual school managed to manipulate the value of the assignment variable in order to be on one side of the threshold rather than the other. If this happens, one might expect to observe a discontinuity in the density of forcing variable at the cut score. Typically, researchers will test the null hypothesis of continuity of the density of the running variable which determines the treatment assignment at the cut score, against the alternative hypothesis of a discontinuity in the density function at that cutoff point (McCrary, 2007). In our case, all schools are evenly distributed on a 100-percentile scale. Hence, by design, it is clearly that there should not be a discontinuity in the density of the percentile ranking at the cut score.

(c) *No jumps at non-discontinuity points assumption*. A third set of specification tests for a zero effect in settings where it is expected that there is no effect (Imbens, 2004). If there is an extraneous discontinuity in dependent variable away from the fixed threshold, the assumption of smoothness in the absence of treatment will be called into question. In practice, we test if the average outcome is discontinuous at other values of the percentile ranking, particularly at the median of the two subsamples on either side of the threshold (Imbens & Lemieux, 2008), where we expect to see no jumps. We choose bottom 2.5% and 7.5% as two placebo cutoff points and test the continuity of school outcomes at each using the specification in equation (1). We find no evidence to reject the null hypothesis of a zero jump at various values of percentile ranking away from the cutoff of 5%. The complete results appear in Table A2.

TABLE A2
Estimated Effects at the Median of the Two Subsamples on Either Side of the Fixed Threshold

	% of students met proficiency level in 2011					
	Reading		Writing		Mathematics	
	z=2.5%	z=7.5%	z=2.5%	z=7.5%	z=2.5%	z=7.5%
Unconditional model	6.332	-4.187	-1.100	0.474	3.199	-7.076
(h=2.5%, n=19)	(9.496)	(11.203)	(9.484)	(10.236)	(3.956)	(7.057)

Note. z = cutoff (z=2.5%: 0-2.5% vs. 2.6-5%; z=7.5%: 5-7.5% vs. 7.6-10%); h = bandwidth; n = sample size. Taken from a separate regression model on schools, each cell in the table shows the estimated coefficient on a dummy variable indicating the effect of being on a placebo list. All estimation models include the forcing variable of percentile ranking as a predictor. Standard errors are in parentheses. Statistical significance is determined using two-tailed tests.
*** p<.001; ** p<.01; * p<.05; †p < .10.

3. PLA List Effects on Percent of Students who were at least Partially Proficient

Similar to our primary estimations looking at the outcomes of the percentage of students who met proficiency level which are major components of school ranking list calculation, all estimated causal impact of PLA list on percent of students who were at least partially proficient in the three subjects are in positive direction, suggesting that schools being placed on the PLA list not only boosts the number of students exceeding the proficiency level but also at the partially proficiency level. This is particularly the case in the ratio of students who were at least partially proficient in math, in which the magnitude and significance level of the subject are substantially larger than those in percent of students met proficiency level. Specifically, being on the PLA list increases the predicted percent of students who were partially proficient in math by 13.8 percentage points in 2011 (p=0.029) and 14.4 percentage points in 2012 (p=0.042).

TABLE A3
Estimated Causal Effects of 2010 PLA List on Percent of Students who were at least Partially Proficient (n=38)

	% of students who were at least partially proficient					
	Reading		Writing		Mathematics	
	2011	2012	2011	2012	2011	2012
Unconditional model	10.960 *	10.031 *	10.866 *	12.488 **	13.821 *	14.364 *
	(6.429)	(5.399)	(4.708)	(4.161)	(7.035)	(8.076)

Note. Taken from a separate regression model on schools, each cell in the table shows the estimated coefficient on a dummy variable indicating the effect of being on the 2010 PLA list. All estimation models include the forcing variable of percentile ranking as a predictor. The covariate variables, including percent of free/reduced lunch students, percent of minority students, school size, and pupil teacher ratio, are collected in the 2009-2010 academic year. Standard errors are in parentheses. Standard errors are in parentheses. Statistical significance is determined using one-tailed tests.
*** p<.001; ** p<.01; * p<.05; †p < .10.

4. Estimated Effects of the 2010 Watch List based on Different Bandwidths

TABLE A4
Estimated Causal Effects of Being on the 2010 Watch List based on Different Bandwidths

	% of students met proficiency level					
	Reading		Writing		Mathematics	
	2011	2012	2011	2012	2011	2012
Unconditional model	0.979	2.505	0.257	5.066	-1.576	2.733
(*h*=10%, *n*=77)	(4.672)	(4.857)	(4.636)	(5.386)	(3.665)	(3.951)
Unconditional model	-2.386	0.197	-5.970	0.605	-7.957	-2.610
(*h*=5%, *n*=40)	(7.151)	(7.749)	(7.183)	(8.077)	(5.228)	(5.651)

Note. *h* = bandwidth; *n* = sample size. Taken from a separate regression model on schools, each cell in the table shows the estimated coefficient on a dummy variable indicating the effect of being on the 2010 PLA list. All estimation models include the forcing variable of percentile ranking as a predictor. The covariate variables, including percent of free/reduced lunch students, percent of minority students, school size, and pupil teacher ratio, are collected in the 2009-2010 academic year. Standard errors are in parentheses. Standard errors are in parentheses. Statistical significance is determined using two-tailed tests.
*** p<.001; ** p<.01; * p<.05; †p < .10.

5. Additional Robustness Testing

(a) *Sensitivity to bandwidth choice.* The robustness and credibility of the primary RD results will further be confirmed if the estimation results are not sensitively rely on a specific choice of bandwidth. To some extent, this sensitivity to the choice of bandwidth test is also one of the RD assumption tests recommended by Imbens and Lemiuex (2008). Table A5 reports the estimated causal impact of being on the PLA list by comparing PLA list schools (bottom 5%) with those control schools 7.5% and 10% above the cut-point. Because of the limited sample size of PLA list schools, we keep the same bandwidth on the left side by including all the treatment schools but alter the bandwidth on the right side to include more non-PLA list schools. As predicted, the magnitudes and significance level of all estimates become weaker and weaker from expanding the group of non-PLA list schools, which ranked higher on the Top-to-Bottom list, from 5% to 7.5% and from 7.5% to 10% above the threshold. Additionally, most of the estimates for the three subjects are smaller in 2012 than those in 2011. Nonetheless, nearly all estimates are in positive direction and some of them achieve significance level particularly in writing.

TABLE A5
Estimated Causal Effects of Being on the 2010 PLA List based on Different Bandwidths

	% of students met proficiency level					
	Reading		Writing		Mathematics	
	2011	2012	2011	2012	2011	2012
-h=5%, n=19; +h=7.5%, n=27						
Unconditional model	6.812	5.740	12.484 *	5.518	2.694	2.041
	(6.093)	(6.395)	(5.682)	(6.812)	(3.392)	(3.545)
-h=5%, n=19; +h=10%, n=38						
Unconditional model	4.889	2.106	7.278 †	3.006	2.127	2.113
	(5.045)	(5.408)	(4.895)	(5.781)	(3.125)	(3.279)

Note. *-h* = bandwidth below cutoff; *+h* = bandwidth above cutoff; *n* = sample size. Taken from a separate regression model on schools, each cell in the table shows the estimated coefficient on a dummy variable indicating the effect of being on the 2010 PLA list. All estimation models include the forcing variable of percentile ranking as a predictor. The covariate variables, including percent of free/reduced lunch students, percent of minority students, school size, and pupil teacher ratio, are collected in the 2009-2010 academic year. Standard errors are in parentheses. Standard errors are in parentheses. Statistical significance is determined using one-tailed tests.
*** p<.001; ** p<.01; * p<.05; †p < .10.

(b) *Small school sample.* In 2010, six regular high schools were too small given the student population to land on the lowest 5% school list. Table A6 shows descriptive statistics of these small schools whereas Table A7 reports the small school comparison results. The upper panel reports the estimated effects of PLA list using all nineteen PLA list schools and six small schools in the regression models. All estimates are in positive direction and some of them reach significance level at 10% given such a limited sample size (25 schools in total). When we restrict our PLA list school sample to those schools which have comparable size as the small schools enrolled fewer than 450 students, all estimates have grown substantially, leading to some of the effects that are significant at the 5% level (as shown in the lower panel in Table A7).

TABLE A6
Descriptive Statistics on PLA and Small Schools

	School Sample	
	PLA (*n*=19)	Small (*n*=6)
	mean (sd)	*mean (sd)*
2011 Outcomes		
% of students met proficiency level in		
Reading	29.42	19.42
	(12.80)	(15.00)
Writing	24.65	14.58
	(13.26)	(14.85)
Math	6.73	2.18
	(5.59)	(4.10)
2010 Pretest		
% of students met proficiency level in		
Reading	29.31	25.52
	(12.10)	(14.47)
Writing	20.46	14.08
	(9.68)	(11.74)
Math	5.87	2.93
	(4.98)	(4.91)

Source: Michigan Merit Examination (MME), Michigan Department of Education (MDE).
Note. n = sample size.

TABLE A7
Effects of Being on the 2010 PLA List vs. Small Schools

	% of students met proficiency level					
	Reading		Writing		Mathematics	
	All PLA schools	*Smallest PLA schools*	*All PLA schools*	*Smallest PLA schools*	*All PLA schools*	*Smallest PLA schools*
	n=19	*n*=6	*n*=19	*n*=6	*n*=19	*n*=6
Unconditional model	9.999 †	18.150 *	10.069 †	14.717 *	4.548 *	7.633 *
	(6.234)	(8.175)	(6.377)	(7.849)	(2.482)	(3.517)
Conditional model (with 2010 pretest)	6.752 *	12.864 *	2.756	6.447	1.923 †	4.496 *
	(3.728)	(5.858)	(3.508)	(5.192)	(1.433)	(1.821)

Note. The number of small schools is six; *n* = sample size of PLA schools. Taken from a separate regression model on schools, each cell in the table shows the estimated coefficient on a dummy variable indicating the effect of being on the 2010 PLA list. The pretest variable is school-level performance in a specific subject in 2010. Standard errors are in parentheses. Standard errors are in parentheses. Statistical significance is determined using one-tailed tests. *** p<.001; ** p<.01; * p<.05; †p < .10.