May 2016

MI

ISSUE BRIEF
Education Policy

GRADING SCHOOLS PROMOTES ACCOUNTABILITY AND IMPROVEMENT

Evidence from New York City, 2013–15

Marcus A. Winters
Senior Fellow

Contents

Acknowledgments

The author thanks participants of the Association for Education Finance and Policy's 41st annual conference for their constructive feedback, as well as the Walton Family Foundation for its generous support.

Executive Summary

In 2007, during the administration of Mayor Michael Bloomberg, the New York City Department of Education (DOE) instituted an accountability system that collected data on NYC public schools and their students' test scores, and issued Progress Reports each fall that included a summary letter grade on each school's performance. In the fall of 2014, Mayor Bill de Blasio replaced the Progress Reports with School Quality Reports and began to change the kind of data that were collected and publicly reported.

Today, NYC schools continue to be assessed on their performance, based on various factors. But the DOE no longer assigns a summary letter grade to each school—nor does it assign any other overall measure of the school's performance. Nevertheless, the data collected and reported in the first School Quality Reports were nearly identical to those collected in the final Progress Reports, issued in the fall of 2013. This means that the summary letter grades for each school that would have been assigned in the fall of 2014—but were not—can be calculated.

This paper explores the effects of the Bloomberg era's school letter grades on NYC's lowest-performing schools; it also estimates the effect of removing these grades after the first year of the new de Blasio accountability system. The paper finds that the decision to stop reporting summary letter grades removed an instrument that had led to positive changes at NYC's lowest-performing schools. **Other key findings include:**

- A positive F-grade impact was detected in the final year (2013) of the original policy, six years after it was first adopted. The effect was smaller than in the first year (2007) but was still meaningful.
- Schools that would have received an F grade in the fall of 2014—the first year of the de Blasio system—showed no improvement relative to schools that would have received higher grades.

I. Introduction

Holding schools accountable for their performance is a central tenet of contemporary education reform. School "report cards"—grading each school with a letter grade from A through F—are a popular accountability tool. Cities and states that use school report cards employ a variety of measures, but all emphasize student performance on standardized tests.

Letter grades offer easily recognizable information to parents and local policymakers about a particular school's quality. It is also widely believed among education researchers that an F grade might help shame a school into improving. Indeed, according to empirical research, that did happen in New York City: schools receiving an F grade made substantial academic improvements the following year, relative to how such schools would have performed had they received a D grade.[1] Research has also shown that Florida schools receiving an F grade made substantial improvements.[2]

New York City graded its schools for the school years 2006–07 through 2012–13, during the administration of Mayor Michael Bloomberg and Schools Chancellor Joel Klein. Shortly after the DOE released its 2013 Progress Reports, Mayor Bill de Blasio and his new schools chancellor, Carmen Fariña, discontinued the practice. In its place, the DOE established a new accountability system that assessed each school's performance on a variety of factors but without any summary measure. Did it matter? Did removing these grades affect these schools' performance?

These questions can be addressed because de Blasio's first School Quality Reports (SQRs), released in the fall of 2014, used the same metrics used in Bloomberg's final 2013 Progress Reports, albeit without a summary letter grade for each school. In this paper, I calculate what these grades would have been. I then use a regression discontinuity design (RDD) that utilizes specific cut points in raw school-performance scores to estimate the causal impact of A–F grades on school performance in the final year of the Bloomberg-era policy—and compare it with what happened in the first year of the de Blasio era. The evidence: schools receiving F grades showed positive improvement, even six years after the grades were first reported. But this improvement dissipated immediately after summary letter grades were dropped.

II. Bloomberg's Progress Reports

Progress Reports included detailed descriptions—and a summary score—of a school's performance in each of several categories: student performance, student progress, educational environment, and bonus points for gains with targeted student populations. The category scores were weighted and combined into a single score to describe the school's overall performance, and the total number of points was translated into a single summary grade between A and F.

These grades were widely reported by local media. The city also warned that schools with low grades for several years would be considered for closure, and many such schools were closed. Over the years, the DOE changed how it calculated the points for each performance category for its Progress Reports, but the categories remained the same.

Two studies evaluated the early impact of NYC's Progress Reports. Rockoff and Turner employed an RDD using school-level data and found that the receipt of an F or an A grade in the fall of 2007 was related to higher average student math and ELA (English Language Arts) test scores in that school the following spring.[3] Winters and Cowen also discovered that schools with an F grade reported higher student test scores, and further found that these effects persisted with the student two years later, suggesting that the scores were not driven by manipulations to the testing process.[4]

Mayor de Blasio's School Quality Reports dropped summary letter grades but continued to provide detailed information about school quality based on various individual factors. While the information in these SQRs has gradually changed,[5] the information in the initial (2014) SQRs was nearly identical to the last (2013) Bloomberg-era Progress Reports. This situation provides a unique opportunity to study the importance of summary grades among very low-performing schools.

Evidence in this paper shows that schools receiving an F grade in the fall 2013 Progress Reports showed gains in their spring 2014 student test scores, relative to other schools. The magnitude of this effect was substantial, though less—as documented in previous research[6]—than the improvement following the first year (2007) of the Progress Reports system. The question is: What happened to the spring 2015 test scores of schools with a (notional) F grade that would have been awarded in the fall of 2014? The improvements in test scores disappeared, as Section III explains.

III. Assessing de Blasio's School Quality Reports

Methodology

The primary data for this paper were acquired from the NYC DOE's website. I supplemented the school accountability reports with school-level aggregate test-score data, also on the website. I employ a version of the RDD to estimate the impact of receiving a particular letter grade on average student test scores within the school the following year. In particular, I apply the basic strategy that Rockoff and Turner[7] used to measure the impact of grades under Progress Reports in the first year after they were adopted. The sample includes all NYC public elementary, middle, and K–8 schools with aggregate test scores, as well as scores under the relevant accountability measure for the year under consideration.

I estimate an OLS (ordinary least squares) regression, where the dependent variable is the school's average test score in the subject under consideration. The independent variables are indicators for the grade that the school received—or would have received—under the grading plan in the fall of that school year. The model controls for an indicator of the school's type (elementary, middle, or K–8), the number of points received by the school on each aspect of the point system, and interactions between school type and the measures of the point system. The regressions include controls for the demographics of students in each school, such as the proportion of students who are members of a minority group, learning English, or in special education.

The "peer index" is an aggregated measure of the school's demographic profile that the city used to identify schools serving similar populations. Rockoff and Turner were able to observe and control for the peer-index score, which had the effect of controlling for a measure of school demographics. The peer index was not made publicly available after 2011; instead, I added other demographic controls. But the results of the regression are similar when I control for the school's peer index on the 2011 Progress Reports.

Previously cited research measured the effect of receiving a letter grade on student performance on math and ELA exams administered in the spring of 2008, following the first letter grades given to schools in the fall of 2007. I expanded this analysis to evaluate whether similar test-score gains continued to be seen in the spring of 2014 (after the fall 2013 school letter grades) and, again, to evaluate what happened to student test scores in the spring of 2015 (after the school letter grades were abandoned in the fall of 2014).

As noted, the first School Quality Reports reported nearly all the data necessary to replicate the summary letter grade that each school would have received in the fall of 2014—had the grade been calculated and publicly reported.[8] I used these data to do just that.[9]

A basic assumption of this paper's RDD procedure is that there are no discontinuities in predetermined characteristics at the cutoffs between grades: since earning points under the accountability system is not directly related to school demographics, we should expect that schools that earned one, or another, letter grade would have similar demographic profiles after we control for the points that it earned under the accountability system.

Figure 1 tests for this attribute by reporting the results from regressions, where the dependent variable is the proportion of students within a school with a particular characteristic, and the independent variables are the schools' letter grade, school type, scores on each component of the point system, an indicator for having a grade that is consistent with the points earned, and interactions between each component of the point system with school level and with the indicator for having a grade consistent with the point total. The results show only two statistically significant differences (and none for those that received an F grade) in demographic characteristics at the grade-level thresholds, suggesting that the assumption likely holds. Or, more simply, there is no statistical difference in the demographic characteristics of those receiving one, or another, grade after controlling for points under the accountability system—thus, the assumption appears to be justified.

FIGURE 1.

Schools with Different Grades Have Similar Demographic Characteristics, 2014 v. 2015

	2014				2015			
	% ELL	% IEP	% Minority	Peer Index 2011	% ELL	% IEP	% Minority	Peer Index 2011
Grade A	0.0159	1.80e-05	-0.0470	-2.748	-0.00105	0.0123	-0.0771	-1.914
	[0.0304]	[0.0154]	[0.0750]	[3.325]	[0.0341]	[0.0167]	[0.0720]	[3.034]
Grade B	0.00327	0.00704	-0.0343	-0.936	-0.00983	0.0101	-0.0980**	-2.788
	[0.0205]	[0.0103]	[0.0503]	[2.182]	[0.0260]	[0.0121]	[0.0469]	[2.006]
Grade C	-0.00254	0.0101	-0.0408	-1.468	-0.0247	0.00586	-0.0677**	-2.308*
	[0.0153]	[0.00798]	[0.0333]	[1.426]	[0.0211]	[0.00953]	[0.0271]	[1.232]
Grade F	-0.0155	-0.00112	-0.0759	-2.107	-0.0144	0.0117	0.0323	1.031
	[0.0244]	[0.0111]	[0.0498]	[2.305]	[0.0302]	[0.0147]	[0.0340]	[1.919]
Observation	1,130	1,130	1,130	1,080	1,157	1,157	1,157	1,091
R-squared	0.195	0.391	0.242	0.801	0.182	0.389	0.284	0.812

Source: Author's calculations using DOE data

Note: ELL = English language learner; IEP = Individual Education Plan (special education); and Minority = a student member of a minority group. Models are estimated via OLS. Robust standard errors are in brackets. The dependent variable is listed at the top of the column. Models include controls for school level, points on each component of the accountability system (progress, performance, environment, and bonus), an indicator for whether the school's letter grade is consistent with total points received, and interactions of points on each component of the accountability system with the school level and with an indicator for whether the grade is consistent with total points received.

* = significant at p < 0.10; ** = significant at p < 0.05

Findings

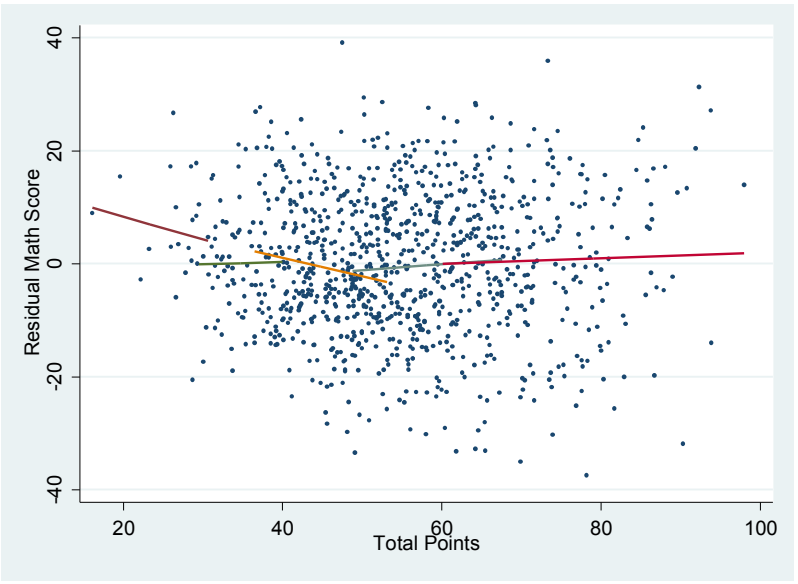
Figure 2 and **Figure 3** plot the residual test scores, after accounting for school level and the components of the point system, against the total number of points that the school received on the accountability system. The lines in the figures represent regressions, where the dependent variable is the school's residual score in math or ELA and the independent variable is total points earned under the system. Each line represents a separate regression run for schools with particular letter grades. For ease of comparison, the figures are restricted to include only schools that had grades consistent with their point total fitting within the bounds of the point total thresholds. The remaining overlap of the lines occurs because of slightly different point cutoffs for elementary, K–8, and middle schools.

Figure 2 shows the jump in performance on the spring 2014 test that occurs for F schools, relative to other schools, following the fall 2013 Progress Reports. This jump is consistent with a positive F-grade effect. Meanwhile, schools that receive other grades appear to perform similarly to one another. Figure 3 shows how schools performed following the 2014 School Quality Reports: schools that would have received an F (had grades been given) perform very similarly to schools that would have received higher grades on the spring 2015 math test.

Because the School Quality Reports system aimed to focus more fully on the school system rather than schools at particular achievement levels, it is possible that its adoption could have altered the entire test-score distribution in a way that would not be apparent-on

FIGURE 2.

School Residual Math Scores on the Spring 2014 Test, Following the 2013 Progress Reports

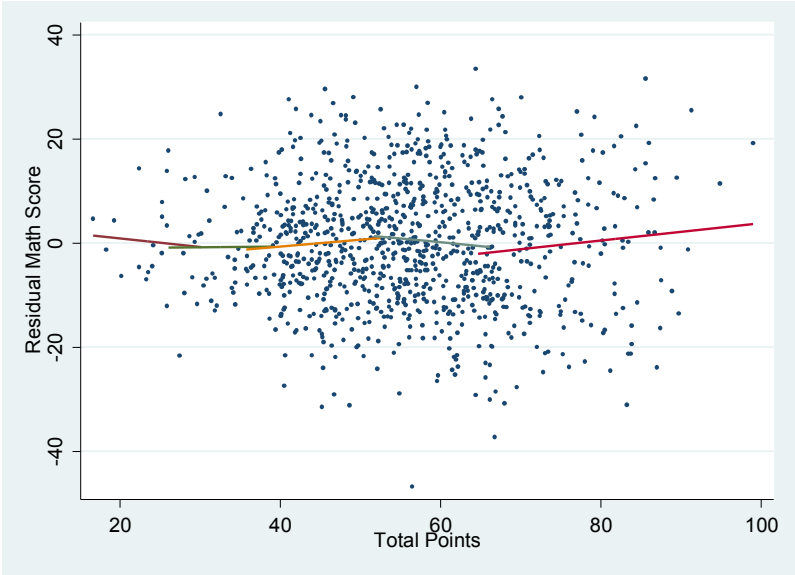


Source: Author's calculations using DOE data

Note: The x-axis is the total points earned under the accountability system, the score upon which the school's letter grade is based. The y-axis is a residual test score—what is left after the effect of demographics and components of the point system is netted out. The residual is what can be attributed to the school's letter grade (or random noise).

FIGURE 3.

School Residual Math Scores on the Spring 2015 Test, Following the 2014 School Quality Reports



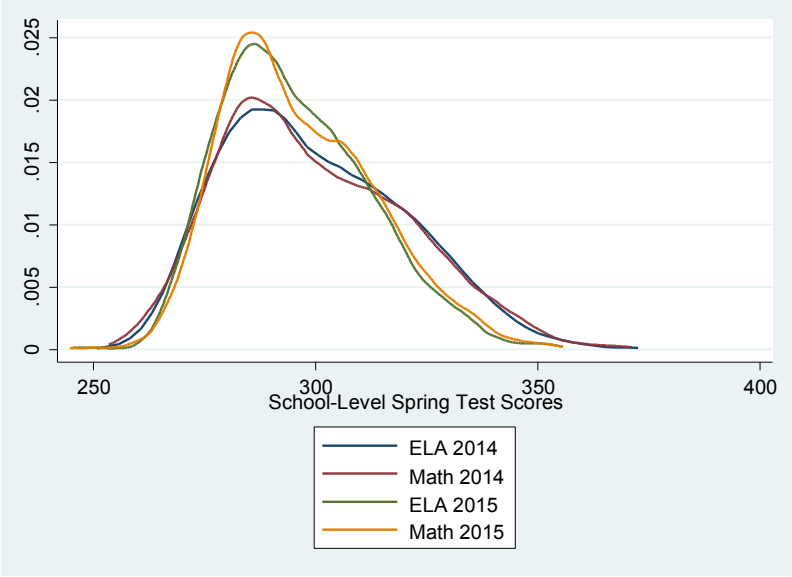
Source: Author's calculations using DOE data

Figures 1–3. To check for this effect, **Figure 4** illustrates the test-score distribution in math and ELA for each of the years under consideration: it does *not* appear that there was a meaningful change in the test-score distribution during 2014–15.

Figure 5 displays the regression-model results. I report results that control for observed school-demographic characteristics—the proportion of students in the school who are black or Hispanic, learning English, or in special education—as of the year under consideration, as well as results that control for the school's 2011 peer index. All models control for the number of points earned in each category, school level, interaction between school level and points in each category, and an indicator for whether the school received a grade other than what would be determined only by the number of points received. In each regression, the comparison group is the test-score improvements in schools that received a D grade.

FIGURE 4.

Test-Score Distributions, 2014 v. 2015



Author's calculations using DOE data

Note: the figure presents a kernel-density function representing the probability distribution for average school scale scores in each New York City school in the estimation sample. The x-axis is the average test score in the school. The y-axis can be thought of as a measure representing the proportion of schools with a particular scale score.

FIGURE 5.

The Effect of Letter Grades on School Test Scores, 2014 v. 2015

	Math				ELA			
	2014	2015	2014	2015	2014	2015	2014	2015
Grade A	-0.777	-3.241	-2.286	-4.172	1.852	-2.802	0.139	-3.287
	[2.024]	[2.582]	[2.705]	[3.222]	[1.827]	[2.148]	[2.381]	[2.537]
Grade B	-0.559	-2.246	-1.731	-2.154	1.368	-1.765	0.198	-1.246
	[1.383]	[1.829]	[1.838]	[2.282]	[1.302]	[1.567]	[1.675]	[1.803]
Grade C	-0.632	-1.923	-1.510	-1.506	1.226	-1.298	0.478	-0.402
	[1.036]	[1.252]	[1.250]	[1.619]	[1.004]	[1.134]	[1.213]	[1.349]
Grade F	3.869*	2.504	4.549*	1.473	3.297**	0.844	3.539**	0.384
	[1.981]	[1.579]	[2.320]	[2.356]	[1.380]	[1.400]	[1.731]	[2.169]
Observations	1,024	1,045	1,024	1,045	1,004	1,015	1,004	1,015
R-squared	0.883	0.894	0.869	0.882	0.788	0.810	0.771	0.806
Current Demographics	yes	yes	no	no	yes	yes	no	no
2011 Peer Index	no	no	yes	yes	no	no	yes	yes

Source: Author's calculations using DOE data

Note: Models are estimated via OLS; robust standard errors are in brackets. The dependent variable is the test score in the spring of the year listed at the top of the column. The letter grade variables indicate either the grade that the school actually received in the fall of 2013 (2014 analysis) or the grade that it would have received in the fall of 2014 had grades been given (2015 analysis). The model includes controls for school level, points on each component of the accountability system (progress, performance, environment, and bonus), an indicator for whether the school's letter grade is consistent with total points received, and interactions of points on each component of the accountability system with the school level and with the indicator for whether the school's grade is consistent with total points received. Current demographics include the percentage of students in the school who are identified as learning English, enrolled in special education, and are black or Hispanic.

* = significant at p < 0.10; ** = significant at p < 0.05

Schools that received an F grade reported significantly improved math and ELA test scores in 2014, following the 2013 Progress Reports, relative to D-graded schools. The magnitude of this difference amounts to about 0.19 standard deviations within the sample in math and 0.17 standard deviations in ELA. To put that result into context, Rockoff and Turner found that receiving an F grade, instead of a D, increased school performance after the first year of the policy by about 0.40 standard deviations.[10] Thus, summary letter grades appear to have a meaningful impact on the performance of F schools six years after these grades were first adopted; but by that point, the impact had been cut in half.

Results from the analysis of 2015 test scores—after the letter grades were removed from the School Quality Reports in the fall of 2014—are consistent with the idea that summary letter grades were driving the impact seen in earlier years. In other words, schools that would have received an F grade in 2014—had it been given—did *not* make improvements, in math or ELA, on the spring 2015 tests relative to schools that would have received a D grade.

IV. Conclusion

Summary letter grades drove improvements in student test scores in New York City schools that received an F grade under Bloomberg's education-accountability system. Students in schools that received an F grade in the fall of 2013 performed better the following year in both math and ELA than they would have, had their school received a higher grade. The magnitude of this difference amounts to about 0.19 standard deviations in math and 0.17 standard deviations in ELA. To put that result into context, Rockoff and Turner found that receiving an F grade, instead of a D, increased student test scores and overall school performance, after the first year of the policy, by about 0.40 standard deviations. Thus, F schools continued to show meaningful improvement even six years after the city's letter-grade policy was adopted, though the magnitude of the effect did appear to drop over time.

Unfortunately, the improvements in student test scores and other performance metrics in F-graded schools ended after the city's Department of Education stopped awarding a summary letter grade, in the fall of 2014. That is, removing the letter grades ended the positive impact on the city's lowest-performing schools.

Since 2014, the School Quality Reports have changed substantially and might soon bear little resemblance to the previous Progress Reports.[11] The new reports have gradually moved focus away from expectations of meeting high standards. Beginning in the fall of 2016, in a shift that perhaps rivals in importance the removal of school letter grades, measures of student progress on standardized tests—the most heavily weighted item under the old Progress Reports system—will no longer appear on the School Quality Reports.

Bloomberg's Progress Reports were not perfect: over the years, the city often tweaked the calculation of points awarded by Progress Reports, as well as how letter grades were given. Nor do the results of this paper suggest that de Blasio's School Quality Reports lack any merit. But this paper does suggest that an effective aspect of New York City's education-accountability policy was lost after school letter grades were dropped. The city should consider reinstating them. Meanwhile, other cities and states that report school grades, but are thinking about ending the practice, should pause to consider the potential consequences for their lowest-performing schools—and the students who attend them.

Endnotes

1 See, e.g., Marcus A. Winters and Joshua M. Cowen, "Grading New York Accountability and Student Proficiency in America's Largest School District," *Educational Evaluation and Policy Analysis* 34, no. 3 (2012): 313–27; and Jonah Rockoff and Lesley J. Turner, "Short-Run Impacts of Accountability on School Quality," *American Economic Journal* 2, no. 4 (November 2010): 119–47.

2 See, e.g., David N. Figlio and Cecilia Elena Rouse, "Do Accountability and Voucher Threats Improve Low-Performing Schools?," National Bureau of Economic Research Working Paper no. 11597 (2005); Rajashri Chakrabarti, "Vouchers, Public School Response, and the Role of Incentives: Evidence from Florida," Federal Reserve Bank of New York, Staff Report no. 306 (2007); Martin R. West and Paul E. Peterson, "The Efficacy of Choice Threats Within School Accountability Systems: Results from Legislatively Induced Experiments," *Economic Journal* 116, no. 510 (2006): C46–C62; and Cecilia Elena Rouse et al., "Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure," *American Economic Journal* 5, no. 2 (May 2013): 251–81.

3 See Rockoff and Turner, "Short-Run Impacts."

4 See Winters and Cowen, "Grading New York Accountability."

5 See Marcus A. Winters, "A Farewell to Reform: NYC's Education-Accountability System," Manhattan Institute Issue Brief 50 (May 2016).

6 See nn. 1–2 above.

7 See Rockoff and Turner, "Short-Run Impacts."

8 Some components in the environment category changed from 2013 to 2014. Thus, I cannot perfectly replicate the previous environmental scores. However, the data elements did not change meaningfully, and the system did present the data necessary to calculate school scores relative to the city overall and relative to a selected peer group. I was therefore able to develop an environmental score using data from the first year (2014) of School Quality Reports that closely replicates the environment score from the final (2013) Progress Reports. The environmental category in the final Progress Reports separated questions on the surveys into the following categories: instructional core, school culture, and structures for improvement. The SQRs collected survey questions in the following categories: rigorous instruction, collaborative teachers, supportive environment, effective school leadership, strong family-community ties, and trust.

9 The grading system is designed so that similar proportions of schools receive each letter grade each year. I began by rank ordering schools of each particular type (elementary, middle, high). I then assigned A grades to the top 25 percent of schools, Bs to the next 35 percent, Cs to the next 30 percent, Ds to the next 7 percent, and Fs to the bottom 3 percent. I then altered grades according to two rules imposed by the system: a school with an average math and ELA proficiency in the top 33 percent citywide can get no lower than a C; and a school that earned an A the prior year can earn no lower than a D. All models included a variable indicating whether the school received a grade outside the bounds of the point system. All models also indicated interactions between the aforementioned variable and each component of the school's score.

10 See Rockoff and Turner, "Short-Run Impacts."

11 See n. 5 above.