

April 2012

Making the Grade?:

Assessing School Progress Reports’ Measurement Of Annual Academic Achievement

Summary

Since the Department of Education presented its first School Progress Reports in 2007, the reports have become one of the primary means for rating and reporting on the effectiveness of each of the city’s schools. Education department administrators use the reports for making decisions about which schools and principals to reward, and conversely, determining which schools to close or principals to remove. Parents and guardians use the reports in helping to choose schools for their children.

Behind each of the letter grades—A, B, C, D, F—characterizing a school’s performance lays a complex measurement system. Given the importance of the annual progress report grades, the education department’s methodology for determining the grades must reflect as accurately as possible a school’s contribution to the annual academic progress of a child, regardless of the child’s prior level of achievement.

Whether the progress reports accurately quantify a school’s contribution to a child’s annual academic progress has been a matter of debate since the reports’ inception. IBO has used sensitivity analyses to examine the education department’s methodology for determining the progress report scores and focused on three key questions: Does the methodology completely control for variables that can affect student outcomes but cannot be controlled by teachers or administrators? Does the methodology capture long-run differences among schools rather than differences that might disappear in a year or two? Does a modest change in methodology lead to substantial changes in the measurement of the data used in the reports?

In general, IBO has found that the methodology used by the education department is a significant improvement over simply basing measures on comparisons of standardized test scores. The use of peer group comparisons reduces biased judgments due to demographics and sampling error between larger and smaller schools. Still, the School Progress Reports have to be interpreted with caution:

- The peer group method may not eliminate all variables that cannot be controlled by teachers and administrators. IBO found that all other things being equal, a school with a higher percentage of black and Hispanic students or special education students is likely to have lower progress report scores.
- The method categorizing schools into letter grades may be sensitive to modest changes. In particular, the distinction between a C and D rating for a school may be the result of a methodological choice by the education department rather than a reflection of school effectiveness.
- Some of the measures used in the progress reports for elementary and middle schools exhibit considerable variation from year to year, although recent changes in methodology by the education department have reduced this volatility. The reports for high schools have always exhibited high levels of year-to-year stability.

Introduction

In the fall of 2007, the New York City Department of Education (DOE) launched a quantitative school accountability system known as the School Progress Reports. These reports are released annually and are meant to be a tool that “enables students, parents, and the public to hold the DOE and its schools accountable for student outcomes and improvement.”¹ They grade each New York City public school along three dimensions: school environment, student performance, and student progress, and then combine them into a single School Progress Report grade. These grades carry practical implications for schools in terms of rewards and consequences. For example, schools that in any given year receive an A or B are eligible for increased funding in the form of bonuses for principals, while schools that receive a D, an F, or a third consecutive C face the possibility of adverse consequences including the dismissal of the principal or even closure of the school. In addition to direct administrative consequences, schools are affected by their progress report scores indirectly, as many parents and guardians use them to inform their own decisions when choosing schools.

Given the important implications of progress report grades, it is essential that the DOE’s methodology for computing them be as successful as possible in fulfilling its goal, which is to “reflect each school’s contribution to student achievement, no matter where each child begins his or her journey to career and college readiness.”² In order to shed some light on the DOE’s success in identifying the contribution schools make to student learning, the Independent Budget Office has analyzed the data and methods used by the DOE to produce the reports for the school years 2006-2007, 2007-2008, 2008-2009, and 2009-2010. As this report was underway, the DOE released the 2010-2011 reports; we have incorporated those results into two of our three research questions below.

This report considers three key research questions on the reliability of these reports for measuring a school’s effectiveness in improving student outcomes:

1. Have the progress reports from 2006-2007 through 2010-2011 for all levels of schools completely controlled for variables that may systematically affect student outcomes but cannot be controlled by teachers or school administrators?
2. Have the progress reports for all levels of schooling

captured differences between schools that persist in the long run, rather than differences that can disappear over the course of just a year or two?

3. Have the progress reports produced estimates that are reasonably robust to modest changes in methodology to how measurements of the same data are performed?

Summary of School Progress Reports Methodology

Goals of the DOE School Progress Reports. School Progress Reports are meant both as a descriptive tool as well as a guide for decisionmakers. In the publicly available *Introduction to the Progress Report and Educator Guides to the School Progress Report*, the DOE specifies the descriptive goal as to provide “an overall assessment of the school’s contribution to student learning,”³ and “produce outcomes that are minimally correlated with socioeconomic status, Special Education populations, or other demographic characteristics.”⁴ The practical goals of the project are stated as follows: “The report is designed to help teachers and principals accelerate academic achievement for all city students. It enables students, parents, and the public to hold the DOE and its schools accountable for student outcomes and improvement.”⁵ This dual purpose implies a degree of trade-off between descriptive accuracy and practical applicability. On the one hand, progress reports must “Measure student outcomes as accurately as possible given the different challenges that schools face;”⁶ on the other, their goal is to “Ensure that schools can verify and re-create metrics so schools understand how they are measured and how they can improve their performance.”⁷

Accurately measuring the contribution of schools to student learning is a task of enormous complexity. Student achievement and progress are affected by a large set of variables; in addition, those variables are nested in a hierarchy of interacting levels (individual, class, and schoolwide). In the presence of such a complicated environment, estimating true school effects based on observational data alone requires very sophisticated (and complex) statistical models. At the same time, the School Progress Report is not an academic exercise; it is meant as a way to give teachers and administrators tools to monitor and improve the performance of their schools. Because of this, the methodology must have a degree of transparency that makes it possible for school managers to anticipate what type of policies could have a positive influence on their students’ education.

DOE Methodology. The procedure for calculating the School Progress Report scores differs across four types of public schools being evaluated—elementary, middle, K-8, and high schools—with the largest methodological differences observed between high school and the remaining three types. For each school, the DOE calculates three separate scores: a school environment score, a student performance score, and a student progress score. Each school’s overall (combined) score is a weighted sum of these three scores.

The school environment score is determined based on responses to surveys filled out by each school’s students, teachers, and parents—there is a separate survey instrument for each group—as well as an analysis of student attendance rates.

The basis for determining student performance and student progress scores varies by type of school. In elementary, middle, and K-8 schools, student performance and progress scores are measured by levels and changes in statewide mathematics and English Language Arts (ELA) examinations. In high schools, student performance is measured by the graduation rate, and student progress by credit accumulation and Regents completion and pass rates.

In school year 2009-2010, a very important modification was introduced with respect to the way student progress scores are calculated for elementary, K-8, and middle schools. Before that year, at the core of the measure were school-level average percentages of students who have achieved a “one year’s worth of progress” with respect to their ELA and mathematics state examination results. A “year of progress” was defined as a student having achieved the same or higher proficiency rating on her ELA or mathematics test in the current year as she did last year, assuming that the current proficiency rating is at least 2.00.⁸ (Proficiency ratings represent the score on the ELA and math test arrayed on a scale ranging from 1.0, 1.1, 1.2 ... to ...4.3, 4.4, 4.5.)

Beginning in 2009-2010, those measures were replaced by school-level median ELA and mathematics student growth percentiles. An individual student’s growth percentile is determined by comparing that student’s current proficiency rating on a given test with current proficiency ratings of all students citywide who one year before had the same proficiency rating as the student in question. A student’s growth percentile indicates the

percentage of students among those who started at the same level, and whose current scores were lower than the student’s.⁹ Growth percentiles are also adjusted to account for differences in certain student demographics, specifically poverty as measured by free lunch status and special education status. Comparing students’ growth percentiles differs from comparing their proficiency ratings in two ways. First, it is independent of scale, which means it will not be affected should there be a noticeable “grade inflation” from one year to the next. Second, by comparing students’ outcomes to those of their peers who were at the same level, it controls for differences in their starting positions.

In addition to student environment, performance and progress scores, schools are also eligible to receive additional points toward their progress report score for Exemplary Student Outcomes, which are awarded based on aggregate measures of unusually large student progress observed among students belonging to particular groups (such as English learner students, special education students, Hispanic and black students in the lowest third in terms of citywide student achievement).

Each of the school-level statistics used to compute a given school’s score on any of the three dimensions are compared with the same statistics of up to 40 schools belonging to its “peer group.” A school’s peer group is a group of schools that serve the population of students most similar to the population of students of the school in question. The similarity of student populations is measured by a one-dimensional “peer index.” Each elementary and K-8 school receives a peer index number between 0 and 100 determined by the percentage of students eligible for free lunch (30 percent of the score) and the percentage of students who are black or Hispanic (30 percent), who are categorized as English Language Learners (10 percent) or who have an Individualized Education Plan, or IEP, (30 percent). Each middle school is assigned an index ranging from 1.0 to 4.5 calculated as the average performance level of currently enrolled students on their fourth grade state exams minus two times the percentage of students with IEPs.

Each high school is assigned an index ranging from 1.0 to 4.5 calculated as the average performance level of currently enrolled students on their eighth grade state exams minus two times the percentage of students enrolled in special education programs minus two times the percentage of students in self-contained classes minus the percentage of students who are over age for their grade. Each school’s peer group consists of 20 schools ranking directly below

it and 20 schools ranking directly above it in terms of the value of their peer index within the group of schools of the same type (elementary, middle, K-8, or high school).

Peer grouping is used to determine 75 percent of a school's progress report scores. The other 25 percent is determined by evaluating each school relative to the citywide range of all schools of the same type. Thus, the progress reports evaluate schools relative to schools with similar demographics, as well as to all schools in the city. (See the sidebar on this page for more details on how peer grouping is used to determine progress reports scores).

The DOE's methodology for translating these peer group and citywide comparisons into letter grades has changed over time. Letter grades for all schools in years 2007-2008 and 2008-2009 and for high schools in 2009-2010 were based on distributions of score cutoff levels, which were

different for each of the four types of schools. In 2009-2010, for elementary, middle, and K-8 schools, the DOE began using a grade distribution based on percentile rankings (previously used in 2006-2007). In order for a school to receive a D, it had to be in at least the 5th percentile of this distribution. Similarly, the threshold to receive a C was the 15th percentile, the threshold for a B was the 40th percentile, and the threshold for an A was the 75th percentile. For high schools, the thresholds were based on score cutoffs and set at 40 points for D, 47 points for C, 58 points for B, and 70 points for A. This basic distribution was later modified with two rules for *ex post* grade adjustment: 1) regardless of its current progress report score and percentile rank, any elementary, middle, or K-8 school that had received an overall grade of A (B) in 2008-2009 could not receive a grade lower than a C (D) in 2009-2010; 2) If a given school's average ELA and mathematics proficiency rating was in the top 25 percent

Use of Peer Grouping in Computing Progress Report Scores

In elementary, K-8, and middle schools, the basis for computing student performance scores are four separate school-level aggregate measures of student achievement: median proficiency level attained in the state mathematics standardized test; that same statistic for the English Language Arts (ELA) test; percent of students who earned a proficiency level 3.00 or higher in mathematics; and the equivalent percentage for ELA proficiency. For each of those basic statistics, its "peer range" and "city range" is then calculated. The city range of a median (ELA or mathematics) proficiency level is the range between the lowest and highest value of that variable observed among all city schools of the same type, excluding extreme outliers.¹ Similarly, peer range of a median proficiency level is the range between its lowest and highest value within a given "peer group" of schools of the same type (again, excluding extreme outliers). Each of the four median scores of every school is then transformed into two new scores, called its "proximity to peer horizon" and "proximity to city horizon," based on where the school's score is placed within both relevant ranges. More precisely, proximity to a group horizon of a school-level statistic *S* in school *i* is calculated as:

$$\text{Proximity to Horizon}(S_i) = (S_i - \text{Reference Group}_i \text{ min}(S)) / (\text{Reference Group}_i \text{ max}(S) - \text{Peer Group}_i \text{ min}(S)),$$

where the reference group is composed of all schools of the same type as *i* in case of city horizon, and of members of school *i*'s peer group in case of peer horizon, and minimum (maximum) of *S* in the reference group of school *i* are defined as the smallest (largest) value of *S* that is not more than two standard deviations away from the group mean. The final progress report score awarded to school *i* for statistic *S* is a weighted average over proximities of *S* to city and peer horizons, multiplied by the maximum score assigned to statistic *S*:

$$\text{Final Score}(S_i) = ((\text{city weight} \times \text{PCH}(S_i)) + (\text{peer weight} \times \text{PPH}(S_i))) \times \text{maximum possible score}(S_i),$$

where *PCH*(*S_i*) and *PPH*(*S_i*) are proximities of *S_i* to their respective city and peer horizons.

The school year 2006-2007 School Progress Reports weighed peer proximities twice as much as city proximities; ever since 2007-2008, peer proximities have been weighed three times as much as city proximities; that is, city weight is equal to 0.25 while peer weight is equal to 0.75. All three component scores of a School Progress Report are calculated from their respective ingredient scores according to the same formula explained above. Total School Progress score of a school is a weighted sum of three basic component scores and (if applicable) the additional credit score.

Endnote

¹Where an "extreme outlier" is defined as a value at least two standard deviations away from the mean of the relevant group.

among all schools of the same type, that school could not receive an overall grade lower than C in 2009-2010.

IBO Analysis of the School Progress Reports Methodology

Controlling for Confounding Variables

Has the DOE methodology completely controlled for variables that can systematically affect student outcomes but cannot be controlled by teachers or school administrators?

In order to adequately compare schools based on the quality of their learning environment one must identify confounding factors—demographic variables that affect student outcomes but are outside the control of the schools—and the prevalence of these factors among schools that are being compared. For example, comparing the English Language Arts test scores of two schools when the student body of one is comprised solely of native English speakers while the other’s includes a large proportion of English Language Learners would unfairly hold the latter responsible for circumstances which it cannot control. The DOE methodology, through its mechanism of peer groups, attempts to control for various demographic characteristics of the student population.

In order to analyze the possible statistical relationship between demographic confounding variables and School Progress Report scores, those scores were treated as a response variable in linear regression models. To test the possibility that progress report scores may be jointly affected by several demographic variables, IBO performed multiple regressions—weighted by enrollment—of student performance, student progress, and overall progress report scores against four school-level demographic variables: percent of students who are black or Hispanic; percent of students enrolled in a special education program; percent of students eligible for a free lunch; and percent of students who are English Language Learners.

This analysis shows that, in elementary schools in every year, the null hypothesis that the joint effect of school-level demographic variables on overall progress report and student performance scores is zero can be rejected at the 0.05 confidence level.¹⁰ In other words, there is no statistical basis to conclude that the DOE’s peer group methodology is completely controlling for the demographic characteristics of students so that student performance scores as well as overall progress report scores for elementary schools are completely neutral with respect to demographics of each

school. The same is true for student progress scores in elementary schools for every year but 2006-2007.

The same applies to overall progress report scores for middle schools, except in 2007-2008.

In contrast, scores for K-8 schools were generally not found to be correlated with the group of demographic variables; the two exceptions were overall and student performance scores in 2010-2011. This means that there are statistical reasons to regard overall progress report scores in K-8 schools as demographically neutral.

All scores for high schools, overall scores, student performance, and student progress scores, for all years, 2006-2007 through 2010-2011, were found to be correlated with the group of demographic variables.

To assess whether progress report scores are correlated with any *specific* demographic variable, IBO performed a series of two-variable regressions of student performance, student progress, and overall scores on each of the four demographic variables separately. All regressions were weighted by school enrollment.¹¹ Weighted Ordinary Least Squares regressions of progress report scores on single demographic statistics are summarized in Table 1 and show the following correlations:

- In all years, 2006-2007 through 2010-2011, high schools’ overall scores and student performance scores are negatively correlated with the percentage of black and Hispanic students, the percentage of special education students, and the percentage of students eligible for free lunch.
- High school student progress scores are also negatively correlated with the percentage of black and Hispanic students and the percentage special education students in all years; and with the percentage free lunch in all years but 2010.
- In 2009-2010 and 2010-2011, in all types of schools, the percentage of students who are black or Hispanic is negatively correlated with student performance, student progress, and overall progress report scores. Those correlations are statistically significant but vary in strength. Overall scores are more highly correlated with the percent of students enrolled in special education than with other demographic variables. The correlations are also stronger in middle schools and high schools,

Table 1: Regression Coefficients Between Progress Report Scores and Demographics, 2006-2007 Through 2010-2011

	Elementary Schools			Middle Schools		
	Overall Score Estimate	Progress Score Estimate	Performance Score Estimate	Overall Score Estimate	Progress Score Estimate	Performance Score Estimate
2006-2007						
Black/Hispanic	-0.058**	0.012	-0.059**	-0.04	0.027	-0.068**
Free Lunch	-0.049*	0.015	-0.062**	-0.002	0.058**	-0.078**
English Language Learner	0.135**	0.086**	0.006	0.129	0.211**	-0.119**
Special Ed	-0.488**	-0.01	-0.369**	-0.256	0.024	-0.19**
2007-2008						
Black/Hispanic	-0.064**	-0.038**	-0.023**	-0.11**	-0.058**	-0.032**
Free Lunch	0.013	0.006	-0.012	-0.057	-0.013	-0.043**
English Language Learner	0.101*	0.055*	0.007	0.055	0.084	-0.074**
Special Ed	-0.108	0.069	-0.165**	-0.075	0.06	-0.107**
2008-2009						
Black/Hispanic	0.003	-0.014	-0.006	-0.047*	-0.048**	-0.002
Free Lunch	0.11**	0.035*	0.007	0.031	-0.012	0
English Language Learner	0.271**	0.155**	0.028**	0.248**	0.142**	0.013
Special Ed	-0.03	0.094	-0.132**	0.165	0.105	0.005
2009-2010						
Black/Hispanic	-0.126**	-0.056**	-0.043**	-0.197**	-0.065**	-0.084**
Free Lunch	-0.033	0.01	-0.035**	-0.057	0.068**	-0.107**
English Language Learner	0.224**	0.164**	0.013	0.221**	0.273**	-0.096**
Special Ed	-0.238	-0.004	-0.211**	-0.25*	-0.024	-0.161**
2010-2011						
Black/Hispanic	-0.135**	-0.064**	-0.042**	-0.214**	-0.081**	-0.075**
Free Lunch	-0.103**	-0.03	-0.054**	-0.203**	-0.026	-0.121**
English Language Learner	0.065	0.052	-0.021	0.01	0.125**	-0.113**
Special Ed	-0.601**	-0.29**	-0.212**	-0.414**	-0.141	-0.176**
	K-8 Schools			High Schools		
2006-2007						
Black/Hispanic	-0.041	0.044	-0.071**	-0.227**	-0.122**	-0.067**
Free Lunch	-0.125*	0.029	-0.121**	-0.122**	-0.062**	-0.059**
English Language Learner	0.265	0.337**	-0.151*	0.026	0.151**	-0.131**
Special Ed	-0.232	0.483**	-0.608**	-1.25**	-0.641**	-0.468**
2007-2008						
Black/Hispanic	-0.074	-0.04	-0.021	-0.171**	-0.09**	-0.05**
Free Lunch	-0.024	0.005	-0.031	-0.106*	-0.06**	-0.036**
English Language Learner	0.186	0.138	-0.061	-0.06	-0.029	-0.06**
Special Ed	-0.092	0.228	-0.302**	-1.01**	-0.561**	-0.305**
2008-2009						
Black/Hispanic	0.001	-0.022	0.008	-0.149**	-0.086**	-0.036**
Free Lunch	0.118**	0.054	0.007	-0.084*	-0.052*	-0.028*
English Language Learner	0.391**	0.241**	0.016	-0.034	-0.033	-0.048**
Special Ed	0.341*	0.406**	-0.16**	-0.866**	-0.526**	-0.265**
2009-2010						
Black/Hispanic	-0.223**	-0.122**	-0.057**	-0.129**	-0.044**	-0.049**
Free Lunch	-0.096	-0.024	-0.066**	-0.077*	-0.018	-0.049**
English Language Learner	0.42**	0.316**	-0.026	0.005	0.003	-0.053**
Special Ed	-0.149	0.168	-0.333**	-0.444**	-0.26**	-0.191**
2010-2011						
Black/Hispanic	-0.164**	-0.076*	-0.055**	-0.163**	-0.086**	-0.047**
Free Lunch	-0.136*	-0.032	-0.088**	-0.101**	-0.045*	-0.047**
English Language Learner	0.111	0.135	-0.076*	-0.031	0.019	-0.062**
Special Ed	-0.457*	-0.065	-0.336**	-0.608**	-0.391**	-0.212**

SOURCE: IBO analysis of Department of Education data

NOTES: *Indicates significance at 0.05 level

**Indicates significance at 0.01 level

where construction of the peer index does not explicitly take student ethnicity into account.

All other things equal, elementary, middle, and high schools with a higher percentage of black and Hispanic students were consistently likely to have lower overall scores than other schools. For elementary schools, each 10 percentage point increase in the proportion of black and Hispanic students generally decreases the school's overall score by more than one point. For middle schools, each 10 percentage point increase in the proportion of black and Hispanic students generally decreases the school's overall score by more than two points. For high schools, the magnitude of this influence is smaller.

Ideally, the progress report methodology would clearly identify the effect that schools have on student performance, independent of the demographic characteristics of the students attending those schools. To date, that goal has not been fully met, though the impact of these demographics on simple measures of school performance (test scores and graduation rates) is known to be much greater than is evident in the progress reports.

There is no statistical reason to conclude that student performance scores as well as overall progress report scores for elementary schools are neutral with respect to demographic characteristics of students. The same results apply to overall scores in middle schools, except for 2007-2008. All scores for high schools overall, student performance, and student progress, for all years, 2006-2007 through 2010-2011 were found to be correlated with the group of demographic variables. In all years, 2006-2007 through 2010-2011, high school overall scores and student performance scores were negatively correlated with the percentage of black and Hispanic students, the percentage of special education students, and the percentage of students eligible for free lunch.

Stability Over Time

Has DOE's methodology captured differences between schools that persist in the long run, rather than differences that can disappear over the course of just one or two years?

Because of frequent changes in methodology, progress report scores or grades are not comparable across years. However, the DOE recognizes that stability over time of progress report measures is a desirable property. For example, when introducing its new methodology of using

student growth percentiles to calculate student progress scores, DOE notes that one of the reasons for this change is that "Growth percentiles will improve the year-to-year stability of the student progress measurement."¹²

Independent of whether or not year-to-year comparisons of progress report measures are valid, their stability over time (in the sense of a "not too large" variance) is an important goal to attain, for both methodological and practical reasons. Assuming that the reported measures adequately capture differences in school quality, low year-to-year stability would imply that the differences captured are transient. If so, the usefulness of progress reports as a tool for developing effective means of improving school quality in the long run would be greatly diminished. In addition, year-to-year stability—even if it is not part of design of the measurement procedure but rather its by-product—is a very useful diagnostic of robustness with respect to nonsystematic disturbances. Given that in a temporally volatile process it is difficult to distinguish signal from noise, then, all else equal, observing such volatility should lower one's confidence that the measure is capturing systematic rather than spurious differences between schools.

Tables 2-5 report Pearson correlation coefficients. Pearson correlation coefficients can take any value between -1 and 1. When increasing values of one variable are accompanied by generally increasing values of the other variable, the coefficient will be positive; it will be negative if increasing values of one variable are on average accompanied by decreasing values of the other variable. A Pearson coefficient close to zero implies little to no correlation between two variables.

Noticeable patterns are:

For schools other than high schools, Pearson correlation coefficients between overall progress report scores from one year to the next have improved with each successive edition. The stability of the overall scores for these schools can be described as weak in the first few years of the reports. Stability of the overall scores jumped with the last two versions of the reports, and can now be described as moderate. (See tables 2 -5)

For this same group of schools, the student progress sub-score is less stable from year to year than either of the other sub-scores and the overall progress report score. The student progress sub-score was very unstable in the early years of the report, but its stability

Table 2: Pearson Correlation Coefficients Between Measures From Different Years of the Elementary School Progress Reports

Measure	2006-2007	2007-2008	2008-2009	2009-2010
	and 2007-2008	and 2008-2009	and 2009-2010	and 2010-2011
Student Performance	0.82	0.84	0.77	0.88
Student Progress	-0.02	0.03	0.21	0.33
School Environment	0.75	0.78	0.82	0.84
Overall School Progress Report	0.21	0.24	0.38	0.53

SOURCE: IBO analysis of Department of Education data

Table 3: Pearson Correlation Coefficients Between Measures From Different Years of the K-8 School Progress Reports

Measure	2006-2007	2007-2008	2008-2009	2009-2010
	and 2007-2008	and 2008-2009	and 2009-2010	and 2010-2011
Student Performance	0.84	0.84	0.72	0.92
Student Progress	0.09	0.06	0.3	0.4
School Environment	0.76	0.78	0.8	0.84
Overall School Progress Report	0.24	0.29	0.47	0.58

SOURCE: IBO analysis of Department of Education data

Table 4: Pearson Correlation Coefficients Between Measures From Different Issues of the Middle School Progress Reports

Measure	2006-2007	2007-2008	2008-2009	2009-2010
	and 2007-2008	and 2008-2009	and 2009-2010	and 2010-2011
Student Performance	0.74	0.69	0.41	0.89
Student Progress	0.07	0.32	0.36	0.43
School Environment	0.7	0.8	0.79	0.84
Overall School Progress Report	0.32	0.47	0.48	0.59

SOURCE: IBO analysis of Department of Education data

has improved in the most recent progress reports, reflecting the stabilizing influence of the student growth percentile metric introduced in 2009-2010 and 2010-2011 reports.

For high schools, correlations from year to year for each of the component scores are generally stronger than those of other school types.

Table 5: Pearson Correlation Coefficients Between Measures from Different Issues of the High School Progress Reports

Measure	2006-2007	2007-2008	2008-2009	2009-2010
	and 2007-2008	and 2008-2009	and 2009-2010	and 2010-2011
Student Performance	0.72	0.74	0.73	0.74
Student Progress	0.66	0.81	0.73	0.69
School Environment	0.83	0.85	0.75	0.76
Overall School Progress Report	0.83	0.84	0.8	0.78

SOURCE: IBO analysis of Department of Education data

For all types of schools, year-to-year correlations of the school environment measure are very strong.

The A-F letter grades assigned to schools for their overall scores have tended to change from year to year (Table 6). Of all the schools that received grades in each of the five years, 53 percent received three or more different grades in the five years, and another 10 percent received four or more different grades. High school grades were more stable than those of other types of schools, with 25 percent receiving three different grades and 3 percent receiving four. The finding that grades for high schools is also supported by looking at the range between the highest and the lowest grade that each school received. For 70 percent of high schools, the range between highest and lowest grade is no more than one level (for example, moving from an A to a B), while only 35 percent of elementary, middle, and K-8 schools had at most a

Table 6: Number of Different Progress Report Grades Received, 2006-2007 Through 2010-2011

Type of School	One	Two	Three	Four	Five	Total
Elementary	28	181	273	62	2	546
K-8	2	38	56	14	0	110
Middle	14	123	101	29	0	267
Subtotal	44	342	430	105	2	923
	5%	37%	47%	11%	0%	100%
High School	48	106	54	6	0	214
	22%	50%	25%	3%	0%	100%
TOTAL-All Schools	92	448	484	111	2	1,137
	8%	39%	43%	10%	0%	100%

SOURCE: IBO analysis of Department of Education data

grade change of one level. Conversely, the range between a school's highest and lowest grade of three levels or more (corresponding to the difference between an A and a D or F) is observed in 7 percent of high schools compared with 19 percent of elementary, middle, and K-8 schools.

All of the five middle schools slated for closure in 2009 have seen their percentile rank improve in 2010, while 10 out of 14 high schools slated for closure in 2009 have improved their percentile rank in 2010. While many schools of all types have improved their percentile ranking from 2008-2009 to 2009-2010, 4 out of 5 middle schools slated for closure have made an improvement larger than most middle schools, and 5 out of 15 high schools slated for closure have made an improvement larger than most high schools (Table 7).

Recent changes in the progress report methodology have made the scores for schools other than high schools more stable from year to year. Those scores can now be described as moderately stable, while earlier editions of the progress reports displayed low stability. The scores for high schools have always been much more stable than those of other schools.

Sensitivity to Measurement Choices

Have the school progress report grades for elementary, middle, and K-8 schools produced estimates that are reasonably robust to modest changes in methodology to how measurements of the same data are performed?

Table 7: 2010 Scores for Schools With Closures Postponed from 2009-2010

School Name	School Type	Progress Report Grade		Overall Score		
		2008-2009 Overall Grade	2009-2010 Overall Grade	2008-2009 Percentile	2009-2010 Percentile	Growth Percentile
P.S. 332 Charles H. Houston	K-8	C	F	1	1	13
Academy of Collaborative Education	M.S.	D	F	1	3	8
Frederick Douglas Academy III Secondary School	M.S.	C	C	2	19	73
KAPPA II	M.S.	D	D	1	10	69
Middle School for Academic and Social Excellence*	M.S.	C	B	4	41	81
New Day Academy	M.S.	C	C	3	22	64
Academy of Environmental Science Secondary High School	H.S.	D	F	4	1	7
Beach Channel High School	H.S.	D	F	1	3	23
Business, Computer Application & Entrepreneurship High School*	H.S.	D	D	2	10	53
Choir Academy of Harlem*	H.S.	D	B	2	42	96
Christopher Columbus High School	H.S.	D	D	2	5	36
Global Enterprise High School	H.S.	C	C	25	19	34
Jamaica High School	H.S.	D	D	4	7	40
Metropolitan Corporate Academy High School	H.S.	D	C	7	11	48
Monore Academy for Business/Law	H.S.	D	C	3	13	62
New Day Academy	H.S.	D	F	6	1	1
Norman Thomas High School	H.S.	D	F	1	2	19
Paul Robeson High School	H.S.	C	C	11	24	74
School for Community Research and Learning	H.S.	C	D	8	6	25
W.H. Maxwell Career and Technical Education High School*	H.S.	D	B	6	31	88

SOURCE: IBO analysis of Department of Education data

NOTE: *These schools were removed from the closure list after the 2009-2010 progress reports were released.

The goal of this section is to determine whether substantively minor changes in the way that particular progress report measures are derived can translate into noticeable changes in School Progress Report rankings. The analysis is limited to the years 2009-2010 and 2010-2011.

Three-quarters of each school's component scores depend on a comparison of that school to a group of its peers. Although differing in a number of ways from the approach used by the DOE, conceptually similar comparisons can be made which are also methodologically appropriate. For example, one can attempt to control for the influence of demographic variables by regressing school-level measures of student outcomes on those variables, and using predicted values as benchmarks against which actual student performance of each school is evaluated.¹³ Tables 8-15 report cross-tabulations of overall progress report grades based on actual 2009-2010 and 2010-2011 progress report scores in each type of school, correlated with overall grades based on scores derived by IBO via quantile regression.¹⁴

The recalculated scores differ from actual scores only in the way in which schools are evaluated relative to their peers. The citywide horizons, the relative weights of city

Table 8: Actual Versus Regression-Based Overall Grades, Elementary Schools, 2010-2011

Actual Grade	Regression-Based Grade				
	A	B	C	D	F
A	0.88	0.11	0.01	0	0
B	0.08	0.79	0.13	0	0
C	0	0.19	0.7	0.11	0
D	0	0	0.27	0.63	0.1
F	0	0	0	0.19	0.81

SOURCE: IBO analysis of Department of Education data

Table 9: Actual Versus Regression-Based Overall Grades, K-8 Schools, 2010-2011

Actual Grade	Regression-Based Grade				
	A	B	C	D	F
A	0.89	0.09	0.03	0	0
B	0.08	0.74	0.16	0.02	0
C	0	0.29	0.63	0.09	0
D	0	0	0.29	0.57	0.14
F	0	0	0	0.29	0.71

SOURCE: IBO analysis of Department of Education data

Table 10: Actual versus Regression-Based Overall Grades, Middle Schools, 2010-2011

Actual Grade	Regression-Based Grade				
	A	B	C	D	F
A	0.87	0.13	0	0	0
B	0.1	0.8	0.11	0	0
C	0	0.16	0.73	0.11	0
D	0	0	0.27	0.58	0.15
F	0	0	0	0.31	0.69

SOURCE: IBO analysis of Department of Education data

Table 11: Actual Versus Regression-Based Overall Grades, High Schools, 2010-2011

Actual Grade	Regression-Based Grade				
	A	B	C	D	F
A	0.79	0.21	0	0	0
B	0.16	0.74	0.09	0.01	0
C	0	0.16	0.75	0.09	0
D	0	0	0.23	0.58	0.19
F	0	0	0	0.36	0.64

SOURCE: IBO analysis of Department of Education data

Table 12: Actual Versus Regression-Based Overall Grades, Elementary Schools, 2009-2010

Actual Grade	Regression-Based Grade				
	A	B	C	D	F
A	0.85	0.15	0	0	0
B	0.1	0.77	0.11	0.01	0
C	0	0.17	0.75	0.07	0.01
D	0	0.02	0.22	0.62	0.15
F	0	0	0.03	0.31	0.66

SOURCE: IBO analysis of Department of Education data

Table 13: Actual Versus Regression-Based Overall Grades, K-8 Schools, 2009-2010

Actual Grade	Regression-Based Grade				
	A	B	C	D	F
A	0.86	0.14	0	0	0
B	0.1	0.71	0.15	0.04	0
C	0	0.27	0.62	0.11	0
D	0	0	0.4	0.47	0.13
F	0	0	0	0.29	0.71

SOURCE: IBO analysis of Department of Education data

Table 14: Actual Versus Regression-Based Overall Grades, Middle Schools, 2009-2010

Actual Grade	Regression-Based Grade				
	A	B	C	D	F
A	0.91	0.09	0	0	0
B	0.07	0.82	0.12	0	0
C	0	0.16	0.69	0.14	0
D	0	0	0.35	0.53	0.12
F	0	0	0	0.24	0.76

SOURCE: IBO analysis of Department of Education data

Table 15: Actual Versus Regression-Based Overall Grades, High Schools, 2009-2010

Actual Grade	Regression-Based Grade				
	A	B	C	D	F
A	0.8	0.18	0.01	0	0
B	0.15	0.76	0.09	0	0.01
C	0	0.14	0.77	0.08	0
D	0	0	0.21	0.7	0.09
F	0	0	0.06	0.18	0.76

SOURCE: IBO analysis of Department of Education data

versus peer horizons, the relative weights of component scores, and additional credit points, are unaltered from the DOE calculations.¹⁵ In order to make the comparison possible, however, letter grades were assigned to both DOE and IBO scores on the basis of percentile ranks.¹⁶ In most of the comparisons, the amount of overlap between actual and regression-based grades is largest for grades A, B, and F, and smallest for grades C and D (with the overlap between schools graded D being considerably smaller than between other grades). This suggests that school-level aggregate statistics may not contain enough information to make a robust distinction between schools that perform at the C level and schools performing at D level.

This exercise highlights a strength of the progress report methodology—its identification of high performing—(A and B rated) schools and very low performing (F rated) schools—stands up to the test of being replicated by a different methodology. A weakness in the progress report methodology is also clearly identified by this test—the distinction between a C and D rating for a school may be the result of the particular methodology that the DOE has chosen, among the many that are possible, rather than the result of school practices or effectiveness. Unfortunately, this weakness occurs at

precisely the point where high stakes decisions about schools are made.

Conclusion

Comparison of performance of schools based solely on their true quality is a difficult undertaking. DOE School Progress Reports are an ambitious attempt at this task. The methodology used by the department is a significant improvement in comparison with accountability methods based solely on standardized test scores in at least two ways.

First, by comparing schools to a set of demographically similar peers, the DOE’s methodology provides a mechanism of controlling for factors that can potentially confound student achievement. This is important, since the lack of such mechanisms can lead to biased judgments of relative school quality.

Second, it provides methods of decreasing bias caused by sampling error. Because of sampling error, grades and schools of smaller size experience much greater variance in standardized test results than schools of average or large size, and are therefore more likely to be punished or rewarded because of factors that are essentially random—a problem with accountability systems based solely on standardized test scores. School Progress Reports correct for sampling error in two ways: by excluding schools that are “far outliers” in terms of results when calculating peer range for school comparisons, and by excluding all items for which there are fewer than 15 observations within any school. In the case of high schools, School Progress Reports provide measurements that are stable over time.

The methodology of School Progress Reports has been evolving over time. The analysis reported here provides evidence that the student growth percentile method of measuring student progress has considerable advantages over the method used previously. It increases year-to-year stability of the measure. Furthermore, the student growth percentile progress measure is also independent of scale, which means that the student progress measure can no longer be affected by possible changes in scale score cutoffs for proficiency levels at the state level. In addition, in 2011-2012 the DOE plans to expand progress reports for elementary and middle schools by adding statistics based on course pass rates, as well as statistics measuring student college readiness for high schools.

This report has performed a number of sensitivity analyses of the DOE School Progress Reports methodology. These analyses provide examples of specific areas in which the outcomes of School Progress Reports have to be interpreted with caution: While the peer group methodology is an improvement compared with other approaches, the method of calculating the continuous metrics on which final progress report scores are based may not fully control for confounding variables. All other things being equal, a school with a higher percentage of black and Hispanic students or special education students is likely to have lower performance and progress scores than other schools.

The method of categorizing schools into letter grades may be sensitive to changes in measuring techniques. There is some evidence suggesting that for overall scores, this sensitivity is largest with respect to distinguishing between C and D grades.

Some School Progress Report metrics display considerable variation over time; the variation is especially large for student progress scores of elementary, K- 8, and middle schools before the introduction of the student growth percentiles method.

This report prepared by Przemyslaw Nowaczyk

Endnotes

¹New York City Department of Education. Educator Guide: The New York City Progress Report, Elementary/Middle/K-8 2010-2011. Updated October 3, 2011.

²ibid.

³ibid.

⁴New York City Department of Education. Introduction to the Progress Report. Updated January 9, 2012.

⁵Educator Guide: The New York City Progress Report, Elementary/Middle/K-8 2010-2011.

⁶Introduction to the Progress Report.

⁷ibid.

⁸After the state ELA and mathematics tests are graded and assigned scaled scores, the New York State Department of Education establishes scale score cutoff levels that place each student in one of the four categories called "proficiency levels," with level-4 being highest. As designed, those proficiency levels are discrete categories; however, for the purposes of calculating student performance scores, DOE transforms them into a continuous measure (called "proficiency rating") based on relative distance of the student's scale score from each of the cutoff values it is placed between.

For example, suppose that a 7th-grader has earned a scale score of 680 on the state mathematics examination, placing herself between the 7th grade cutoff scores for mathematics performance levels 3 and 4 which are 670 and 693, respectively. This student's mathematics proficiency rating is $3 + ((680 - 670) / (693 - 670)) = 3.43$. Students who exceed the cut-off score for performance level 4 are awarded proficiency ratings between 4.01 and 4.50.

⁹ For example, to calculate the growth percentile of a student that has earned an ELA proficiency rating of 3.1 in his 3rd grade examination in 2009, and 3.4

in his 4th grade exam in 2010, this student's 4th grade result is compared to 2010 results of all 4th-graders citywide who scored 3.1 on their 3rd grade ELA test. If, say, 57 percent of students in this comparison group scored lower than the student in question, then that student's ELA growth percentile is 57.

¹⁰Via the F-test of joint significance.

¹¹In several unweighted OLS regressions, the Koenker-Basset test for heteroskedasticity rejects the null hypothesis of homoskedasticity.

¹²New York City Department of Education. Final Changes to Elementary and Middle School Progress Reports 2009-2010. Updated September 27, 2010.

¹³Chay, K. Y., McEwan, P. J., and Urquiola, M. (2005). The central role of noise in evaluating interventions that use test scores to rank schools. NBER Working Paper No. w10118.

Kane, T., and Steiger, D. O. (2002a). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives*. 16:4, pp. 91-114.

Kane, T., and Steiger, D. O. (2002b). Improving school accountability measures. NBER Working Paper No. 8156.

and Hallock, K. (2001). Quantile Regression. *Journal of Economic Perspectives*. 15:4, pp. 143-156.

¹⁵For each school-level aggregate statistic among schools of the same type, IBO has used quantile regression to estimate each percentile of the distribution of that statistic conditional on school-level demographic variables. Each school's peer horizon score is substituted by its predicted percentile (that is, percentile in which residual is smallest), as estimated by those regressions. For each type of schools, the demographic variables used in those regressions are the same variables that are used by DOE to construct peer indices.

¹⁶With 25 percent of schools assigned A, 35 percent B, 25 percent C, 10 percent D and 5 percent F.

Receive free reports by e-mail
iboenews@ibo.nyc.ny.us
Twitter: twitter.com/nycibo
RSS: www.ibo.nyc.ny.us/iborss.xml