Estimating Blue Engine's Impact on Students' Performance on State Standardized Tests: An Overview of Methods and Results

AY 2012-2013

November 2013

Rebecca Casciano
Glass Frog Solutions
rebecca@glassfrog.us

Dawn Perlner
Glass Frog Solutions
dawn@glassfrog.us

This report was completed in partnership with Glass Frog Solutions.

P.O. Box 9
Skillman, NJ 08558
1.888.609.3372
info@glassfrog.us
www.glassfrog.us

This report presents results from a project conducted by Blue Engine, a nonprofit 501(c)(3) charitable organization that places teaching assistants (BETAs) in classrooms at New York City partner high schools. BETAs work alongside math and English teachers within these schools and also assist in implementing a curriculum designed by Blue Engine to teach social cognitive skills to students. The overarching goal of the program is to help students be admitted to, enroll, and graduate on time from college.

The goal of the project is to develop a methodology for using previous state standardized test data from the New York City Department of Education to forecast[1] performance on these tests among current and future Blue Engine students. These forecasts have a dual purpose: (1) to enable Blue Engine to set goals for student performance that are rigorous yet realistic because they are rooted in knowledge of how similar students have performed historically and (2) to evaluate the program's impact on performance with statistical evidence of improvement in scores due to the program. This report explains the forecasting methodology in detail and provides an analysis of how Blue Engine students performed relative to their predicted scores and relative to Blue Engine's internal goals. It concludes with recommendations for improving the forecasts in future years.

Measuring student performance at Blue Engine

Blue Engine aims to prepare high school students for college, so it requires a reliable way to gauge whether students are improving on measures of college readiness. The organization views performance on standardized tests as one indicator of the extent to which students have mastered course material, and therefore judges its success in part based on students' exam scores.

New York State uses statewide tests called Regents exams that assess knowledge of core high school subjects. To graduate from high school, a student must pass exams in algebra, English, science, and social studies. Since Blue Engine only works with students in math and literacy, it uses students' scores on the Integrated Algebra, Geometry, Algebra 2, and English Language Arts (ELA) Regents exams as barometers of its impact.[2]

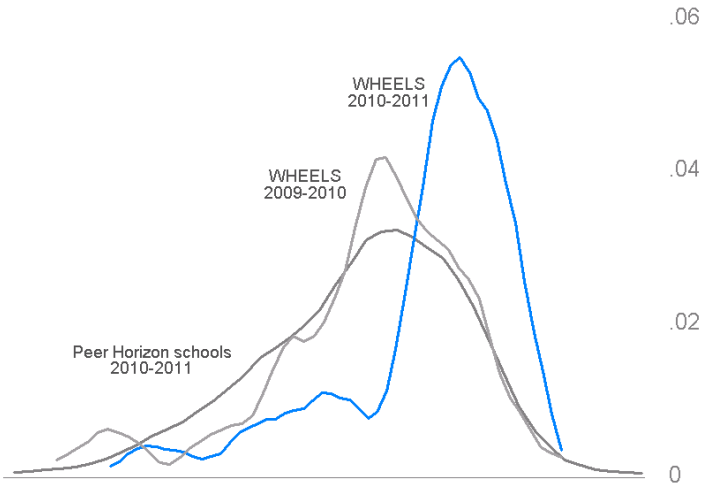
In the past, Blue Engine has measured its impact on student performance on Regents exams in several ways. It examined differences in average scores between Blue Engine students and specific comparison groups, such as previous cohorts of students within the same schools or students attending similar (non-Blue Engine) schools in New York City.

Average scores offer a simple and easy-to-understand metric of student performance; however, we have and can learn from more information than this. For example, the distribution of scores may reveal that certain groups of students perform better or worse than their peers or, in the case of Blue Engine students, benefit more or less from the program. To give a simple

[1] We use the terms "predict" and "forecast" interchangeably in this report.

[2] This report presents results from Algebra, Geometry, and English Language Arts, but will update the findings to include results from Algebra 2 once we have access to the data.

Figure 1: Example of density plots comparing Blue Engine student performance to performance among students in relevant comparison groups.



example, in a class with ten students, five of which are in Group H and five in Group L, the H students may all score 80 points, and the L students all score 60, giving the class an average of 70. However, closer examination of scores tells us that the students in Group L may need more help. The Blue Engine program, if not effective across categories, may raise the H students' scores to 100 while leaving the L students' scores at 60. The class average has now gone up to 80, and examination of the averages tells us that Blue Engine has been effective at raising students' scores, even though it has not been effective at raising the L students' scores. Moreover, if a score of 65 is required to pass the exam, the passing rate has not improved at all: the average student's score has gone from failing to passing, but no actual students have passed the exam who would not have passed before.

Thus, Blue Engine also examines differences in the *distribution* of scores relative to these comparison groups. This exercise is typically accomplished by creating density plots of student scores and examining differences between Blue Engine and non-Blue Engine students in the location and shape of the plots. An example of this exercise is shown in Figure 1.

Additionally, Blue Engine is particularly interested in whether students score above specific thresholds that indicate satisfactory or advanced mastery of course material. The first threshold indicates whether students pass a Regents exam with a score of 65 or higher; this is required for students wishing to obtain a Regents diploma in New York State.[3] The second threshold indicates whether students score at or above a predefined college "ready" threshold. The college ready thresholds were determined by a Harvard University study that followed the academic progress of New York students studying at the City University of New York. The authors found that, in order to receive a C average in entry level course work, students needed to score a minimum of an 80 on the algebra Regents exam and a 75 on the ELA Regents exam.[4] Each year, Blue Engine computes the portion of students who pass and score

[3]Students with Individualized Education Programs pass the core subject exams with a score of 55 or higher.
[4]More background on the origins of these thresholds can be found here:

college ready on each exam.

As mentioned, Blue Engine has experimented with using various comparison groups against which it can measure the progress of its own students in a non-experimental context. These comparison groups include students attending a Blue Engine school before Blue Engine partnered with the school, students attending comparable schools in New York City, and students citywide. Each group has its strengths and weaknesses as a comparison group. For instance, students working with Blue Engine might be somewhat similar to previous cohorts within the same school who did not work with Blue Engine; however, the comparison is rarely ideal. For example, since Blue Engine aims to accelerate the rate at which students learn course material and take Regents exams, they typically encourage all students in a grade to take an exam (e.g., all ninth graders will take the algebra exam, all tenth graders will take the geometry exam, etc.). However, sometimes it is the case that only a subset of students in the previous cohort (i.e., before Blue Engine entered the school) took an exam in the prior year, and typically this subset is higher achieving than the full cohort of students would be (e.g., only the top 30 students might take the ELA exam in the tenth grade). In these cases, Blue Engine must compare the performance of a full class of students to the performance of a select group of students from the previous cohort. Moreover, the longer Blue Engine works in a school, the less relevant this "pre-Blue Engine" cohort is as a comparison group, since there are likely other factors changing over time in these schools.

Blue Engine can also compare its students to students attending similar schools in New York City. This method has the advantage of providing a comparison group of students taking the same test in the same year as Blue Engine students (i.e., not a test from a prior year, as would be the case if we were using previous cohorts as the comparison group). However, using students from other schools as a comparison may introduce selection bias insofar that these schools may differ in unmeasurable ways from schools that partner with Blue Engine. For instance, it is possible that Blue Engine schools have more innovative leadership teams, on average, than non-Blue Engine schools and that innovative leadership can explain both their partnership with Blue Engine *and* and their students' strong performance on state exams.

This report describes an alternative method of gauging Blue Engine's impact on student performance. The method enables Blue Engine to measure impact using each of the metrics described above: average scores, distributions, and portion passing/college ready. However, instead of measuring impact against an external comparison group, it compares students' actual performance to how they are predicted to score based on how similar students from previous cohorts have performed.

This method can be used to measure Blue Engine's impact after students have taken their exams, but it can also be used to set realistic yet rigorous internal goals long before students take the exams. This is an innovative use of data for Blue Engine, which previously had to rely on a mix of intuition and assumptions about how their students would perform based on prior experiences with students. Moreover, by providing an alternative for modeling the counterfactual (i.e., how these students would have performed in Blue Engine's absence),

http://gothamschools.org/2011/02/11/college-readiness-may-take-even-more-than-states-stats-show/.

this predictive approach raises the organization's standards for assessing its impact, aligns the organization more closely to the standards of leading impact funders like The Robin Hood Foundation, and helps the organization prepare for more formal, high-stakes impact evaluations.

Modeling student performance on state standardized tests

In the 2012-2013 academic year, Blue Engine partnered with three schools in upper Manhattan and the Bronx. We refer to those schools as School 1, School 2, and School 3 in this report. To predict Regents scores for students working with Blue Engine in these three schools, we draw on data from previous cohorts of test takers in the New York City public school district. These data are available by request from the New York City Department of Education.

A central goal of the project is to predict how Blue Engine students would perform on Regents exams in the absence of the Blue Engine intervention. Thus, for each subject and school, we based the predictions on how comparable students across the district performed on the test in the academic year prior to Blue Engine's first year working in a school. For example, Blue Engine began working with algebra students at School 1 in the 2010-2011 academic year; we therefore based the predictions for School 1 algebra students on citywide data from the 2009-2010 academic year. Similarly, Blue Engine began working with English Language Arts students at School 2 in the 2012-2013 academic year, so we used citywide data from the 2011-2012 year for the predictions. We used this methodology because, as we describe in more detail below, we generated school fixed effects estimates and therefore needed to use "pre-Blue Engine" data for the models, or else the school effect would also reflect the Blue Engine effect. Table 1 shows the years in which Blue Engine began working with each subject within each school.

For each subject, we generated predicted scores using a multivariate regression model, a statistical method used to study the association between a dependent variable (in this case, the Regents exam score for a particular subject) and one or more independent variables (e.g., other test scores, special education status, etc.). When there is more than one independent variable in the model, the association between one independent variable, X, and the dependent variable, Y, is computed by holding the other variables in the model constant and then estimating the relationship between X and Y: that is, the slope of the line that would result from a graph of Y on X.[5]

In the context of the social sciences, multiple regression can be used to explain behavior and trends: for example, the association between the stringency of drug laws and the incidence of drug-related crime across U.S. cities or the role that maternal education plays in determining childhood nutrition. However, multiple regression also offers a method of using existing data

[5]Obviously the actual data points do not usually fall exactly on the line; this is due to random error that may be caused by hidden variables we cannot measure, measurement error, or other factors. We employ post-estimation measures to gauge whether our errors are actually random and unassociated; that is, that the line we've created is a good predictor of our data and not biased in some way.

Table 1: Schools and subjects working with Blue Engine, by year.

School	2010-2011	2011-2012	2012-2013
School 1			
Algebra	X	X	X
Geometry	-	X	X
ELA	-	X	X
School 2			
Algebra	-	X	X
Geometry	-	-	X
ELA	-	-	X
School 3			
Algebra	-	-	X
Geometry	-	-	X
ELA	-	-	-

to create a formula that can then be used to predict (or "forecast") future behavior or trends. This is the reason for using multiple regression in the present paper. We draw on student test score data from previous years to create a formula that is used to predict the performance of future students.

Specifically, we predicted students' scores on a particular Regents exam based on their previous performance on state standardized tests, whether they were classified as English Language Learners (ELL), and whether they had an Individualized Education Program (IEP) (i.e., are classified as requiring special education). Table 2 shows averages on these characteristics for Blue Engine students and for the populations of students on which the forecasts are based. With some minor variation, Blue Engine students' eighth grade scores and IA Regents scores were fairly comparable to those of students in the model populations. However, Blue Engine students had a somewhat greater share of students with IEPs and a much greater share of students classified as ELL.

We accounted for the fact that students in New York City attend schools with variable quality and resources by estimating school "fixed effects," which is tantamount to including controls for a student's school in the model.[6] By incorporating school fixed effects, we controlled for

[6] In a fixed effects model, dummy variables are created representing all possible values of a dependent variable which is constant among subsets of the data. In this case, we created a fixed effect for school: each student was attributed to one of the participating schools, and each school included multiple students. Therefore, if a student attended School 1, her School 1 dummy would be one, and her dummy statistic for all other schools would be zero. To estimate the fixed effects models, we used Stata's `areg` command, which automatically creates the dummy variables for all represented values of the specified variable (in this case school). Stata

Table 2: Comparing Blue Engine students with student populations used to generate forecast weights (BE=Blue Engine; Model = Model population).

	8th grade math (mean)		8th grade ELA (mean)		IEP (%)		ELL (%)	
	BE	Model	BE	Model	BE	Model	BE	Model
Algebra								
School 1	671	647	–	–	25	10	27	12
School 2	665	659	–	–	21	18	18	9
School 3	668	667	–	–	23	17	18	9
Geometry								
School 1	682	682	–	–	18	6	11	4
School 2	674	687	–	–	15	7	6	3
School 3	665	687	–	–	7	7	8	3
ELA								
School 1	–	–	648	647	18	13	10	6
School 2	–	–	649	655	12	13	7	5

everything about a school that might impact student achievement, including school climate, teacher quality, leadership, school culture, etc. Including the school effects improves the models because it allows us to measure a school's average contribution to student test scores, above and beyond a student's own individual characteristics. In some cases, a school effect is negative (e.g., -3.1), which means that students' predicted scores in that school should be adjusted downward by that fixed amount (i.e., by the size of the effect). Conversely, in a school where the fixed effect is positive (e.g., + 1.9), all students' predicted scores would be adjusted upward by that amount.

It is important to account for these school fixed effects because relying solely on students' individual characteristics to predict scores can lead to systematically low (or high) estimates in a given school. For instance, suppose Blue Engine partners with a school in 2013-2014 where students scored very high on the ELA Regents exam in the previous year (2012-2013). Now suppose that students in this previous cohort were not exceptional in any other way: they had average scores on their eighth grade exams and average IEP and ELL rates. If we only relied on students' background characteristics to generate the predictions, then these

uses two dummy variables at a time (in the set, not in the set), calculating the fixed effect for one possible value of the variable in question at a time. For example, while calculating school fixed effects, Stata will run the OLS regression with enrollment in each school, and the set of all other schools, as the two extra dummy variables. By rotating through all possible schools, a fixed effect is obtained for each school. The fixed effect represents the effect of just being part of the specified subset of data: for example, being at School 1 may add, on average, three points to a student's Regents score. Therefore the fixed effect for School 1 would be three. Without a fixed effects model, the school effect would become part of the random error for each student; the fixed effect allows us to attribute part of that random error to a particular cause (in this case, what school the student is in), and therefore improve the fit of our model, if in fact expected scores differ across schools.

students' predicted scores would tend to be very average. They would not reflect the high overall ELA performance of students in this school. At the end of Blue Engine's first year working in this school, it would likely discover that students performed much higher than predicted and may erroneously take credit for at least some of this feat. If the forecasting models had accounted for school fixed effects, the predictions would have been much higher, which would have produced a more conservative estimation of Blue Engine's "effect" in this school.

For comparison purposes, we estimated the same models without school fixed effects. Appendix A shows the portion of students predicted to pass and score college ready based on the non-fixed effects models.

The equation can be written as follows:

$$REGENTS_{ij} = \beta_0 + \alpha_j + \delta PREVIOUS_{ij} + \beta_1 ELL_{ij} + \beta_2 IEP_{ij} + \varepsilon_{ij}$$

where $REGENTS_{ij}$ is the Regents score of student i in school j, α_j is the school fixed effect for school j, and ELL_{ij} and IEP_{ij} are binary variables equal to one if the student has, respectively, ELL or IEP classification. $PREVIOUS_{ij}$ is a vector of controls for previous performance on state standardized tests. For the algebra and geometry models, this includes the student's score on the eighth grade state math exam. For the ELA model, it includes the student's score on the eighth grade state ELA exam. In order to improve model fit, we included a quadratic term on the student's previous exam scores, as well as dummy variables indicating whether a student scored either very high or very low on these exams. (We included these terms because the models tend to under-predict Regents scores for students who scored very low on previous exams and, to a lesser extent, over-predict scores for students who scored very high on previous exams.) The coefficients from these models are shown in Table 3. The models fit the data fairly well, explaining between 44% and 63% of the variance in students' scores on the Regents exams, depending on the exam (see the R-squared row in Table 3).

Forecasting performance for current students

Once we estimated this equation using data from prior years, we used the coefficients in Table 3 as a formula to forecast scores for all Blue Engine students. As explained above, the predicted scores are generated from a multiple regression model. They can therefore be interpreted as the Regents score a student can expect to obtain given his/her previous test score(s), IEP status, and ELL status. For example, "John Smith" is an algebra student at School 3. He scored 670 on his eighth grade math exam, does not have ELL status, but has an IEP. The above formula would yield a predicted score of 62.5 on the Integrated Algebra Regents exam for John. This was computed using the following formula:

SCORE = 322.227 + -.833 + 670*-1.029 + (670^2)*.001 + 0*-10.316 + 0*3.700 + 1*-4.950 + 0*-1.906

Table 3: Coefficients from ordinary least squares regression models estimating New York City school students' scores on Regents exams.

	Algebra			Geometry		ELA	
Variables	09-10	10-11	11-12	10-11	11-12	10-11	11-12
Constant	231.934	286.293	322.227	-18.673	-251.274	-170.594	-600.735
8th grade math score	-0.703	-0.883	-1.029	.480	0.612	–	–
8th grade math score^2	0.001	0.001	0.001	-0.000	-0.000	–	–
8th grade ELA score	–	–	–	–	–	0.535	1.740
8th grade ELA score^2	–	–	–	–	–	-0.000	-0.001
8th grade score: high	-9.725	-11.700	-10.316	-5.542	-6.056	-7.401	-7.285
8th grade score: low	2.412	1.739	3.700	16.599	31.944	13.465	70.425
IEP	-7.208	-5.194	-4.950	-3.746	-4.343	-7.160	-7.666
ELL	-1.419	-1.786	-1.906	-0.381	-0.488	-3.112	-3.449
R-squared	0.436	0.476	0.536	0.596	0.627	0.478	0.522
N	77815	76518	54464	40808	40593	73670	60612
School Effect							
School 1	2.748	–	–	4.316	–	2.808	–
School 2	–	-1.730	–	–	-17.888	–	2.394
School 3	–	–	-.833	–	-3.399	–	–

Since we are using scores from previous years to predict the performance of students in the 2012-2013 academic year, a central assumption is that the predictors (i.e., previous scores, ELL status, IEP status) will be related to Regents scores in the 2012-2013 year in the same way as in previous years. We cannot know if this is true until the 2012-2013 citywide data are released by the New York City Department of Education. In the meantime, to test the effectiveness of using data from year n to predict scores in year $n + 1$, we used the district-wide data available to us from one year (e.g., 2010) to predict scores among students in the successive year (e.g., 2011). We then compared our predictions against those students' actual scores.

Table 4 shows the portion of cases our models correctly predicted to pass or fail (or score college ready/not score college ready). The predictions of college readiness tend to be more accurate than the predictions of pass rates. This is in part a simple matter of probability. Most students have a good chance of passing the exam, as the threshold is fairly low. Therefore, a large portion of students will pass or fail according to random variation: how they feel on the day of the exam, how much they studied for it, and other things we cannot measure. There are a few students we can be almost sure will pass, or almost sure will fail, but for the most part, our model will not capture the unobserved life circumstances that swing students who are on the verge of passing one way or another. Therefore, our predictions for a large percentage of the students will randomly be accurate or not as the student randomly does

Table 4: Portion (%) of cases correctly predicted to pass (or fail) and score college ready (or not score college ready). The estimates reflect predictions for students from 2011-2012 based on coefficients generated using data from 2010-2011 students.

	Passing/failing	College ready/not college ready
Algebra	75.2	87.2
Geometry	78.5	82.4
ELA	83.6	74.9

well or poorly on the day of the exam.

For college readiness, we are also subject to these sorts of random variations, but as the bar is set so high, very few students' college readiness will be affected by those variations; most students simply will not score college ready, even on a good day. We will predict correctly the majority of students who have no hope of scoring college ready, and will randomly predict correctly for the rest. Therefore our percentage accuracy is bolstered by the large portion of students we are sure will not score college ready. Note that for ELA, the standard for college readiness is five points lower, and accordingly, a larger percentage of students have a chance at meeting the goal, and our predictions are therefore slightly less accurate.

There is no obvious pattern as to whether failed predictions are too high or too low; further study is required. For example, it is possible that students may not spend much effort preparing for their eighth grade state exams and then take the Regents exams more seriously because they have more bearing on their futures, or simply because they are more mature by the time they take them. However, there are false positives (students incorrectly predicted to pass) as well. In the Recommendations section, we discuss how using data from additional years as well as working with more complete data for Blue Engine students might improve the predictions.

Notably, while Blue Engine aims for all of its students to take the Regents exams, some students were absent on the day of the exam. The absence rate varied across school and subject. Whereas only 2.5 percent of all algebra students and 3.1 percent of all ELA students were absent, 19.5 percent of geometry students did not show up for the spring exam. A full accounting of missing data can be found in Appendix B.

Using predicted scores to gauge impact

As mentioned earlier, one purpose of generating the predicted scores is to help Blue Engine assess its impact in a non-experimental context. To do this, one must only compare the students' predicted scores to their actual scores on the exams. Table 5 shows mean predicted scores as well as the portion of students predicted to pass and score college ready alongside students' actual performance. Figure 2 presents the number of students not passing, passing,

Table 5: Comparing predicted versus actual scores among Blue Engine students.

	Mean score			% Passing			% College ready			
	Predicted	Actual		Predicted	Actual		Predicted	Actual		N
Algebra										
School 1	69.4	71.0		100.0	79.7	*	**0.0**	**23.7**	*	59
School 2	64.8	63.2		66.3	57.1		**0.0**	**9.2**	*	98
School 3	**66.1**	**74.1**	*	**70.9**	**90.9**	*	**1.8**	**27.3**	*	110
All	**66.3**	**69.4**	*	75.7	76.0		**0.8**	**19.9**	*	267
Geometry										
School 1	72.6	67.7	*	90.3	64.5	*	**6.5**	**29.0**	*	62
School 2	**50.6**	**62.7**	*	**0.0**	**48.7**	*	**0.0**	**5.3**	*	76
School 3	**59.2**	**68.2**	*	**19.0**	**73.0**	*	**1.0**	**12.0**	*	100
All	**59.9**	**66.4**	*	**31.7**	**63.0**	*	**2.1**	**14.3**	*	238
ELA										
School 1	76.3	76.9		100.0	96.1		77.6	71.1		76
School 2	74.4	71.6	*	97.4	87.0	*	55.8	44.2	*	77
All	75.3	74.3		98.7	91.5	*	66.7	57.5		153

Note: Statistically significant ($p<.05$) differences between predicted and actual scores (or predicted and actual proportions) are denoted by a *. Instances where students' actual scores are significantly greater than predicted scores are shown in bold.

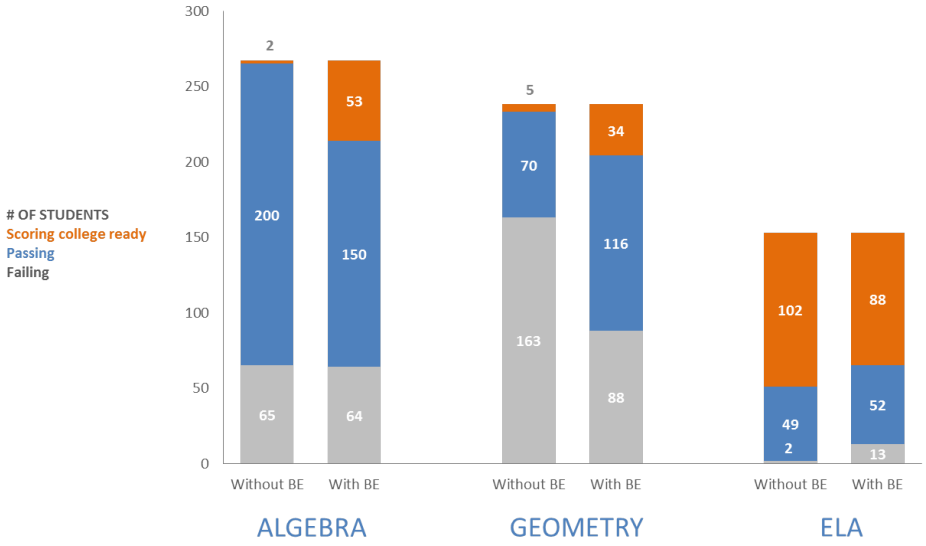
and scoring college ready across the three subjects, aggregated across schools.

Overall, Blue Engine students scored better than predicted on the Integrated Algebra Regents exam. A comparable proportion of students passed the exam relative to what was predicted (predicted and actual pass rates were both 76%), but the portion scoring college ready exceeded the portion predicted to score college ready by 19 percentage points. Average scores were roughly three points higher than predicted.

There was considerable variation across schools on the algebra exam. At School 1 and School 2, students' scores on the algebra exam were comparable to predicted scores and smaller portions of students passed the exam than predicted. However, the percent scoring at a college ready level was higher than expected in both schools: 23.7 percentage points higher at School 1 and 9.2 points higher at School 2. Algebra students at School 3 scored considerably higher than predicted: 90.9% passed the exam (70.9% predicted) and 27.3% scored at a college ready level (1.8% predicted).

In general, students also scored better than predicted on the Geometry Regents exam. Average scores were 6.5 points higher than predicted, and the portions passing and scoring college ready were 31 and 12 points greater than predicted, respectively. Once again, performance varied

Figure 2: Number of Blue Engine students not passing, passing, and scoring college ready on Regents exams: Predicted versus actual.



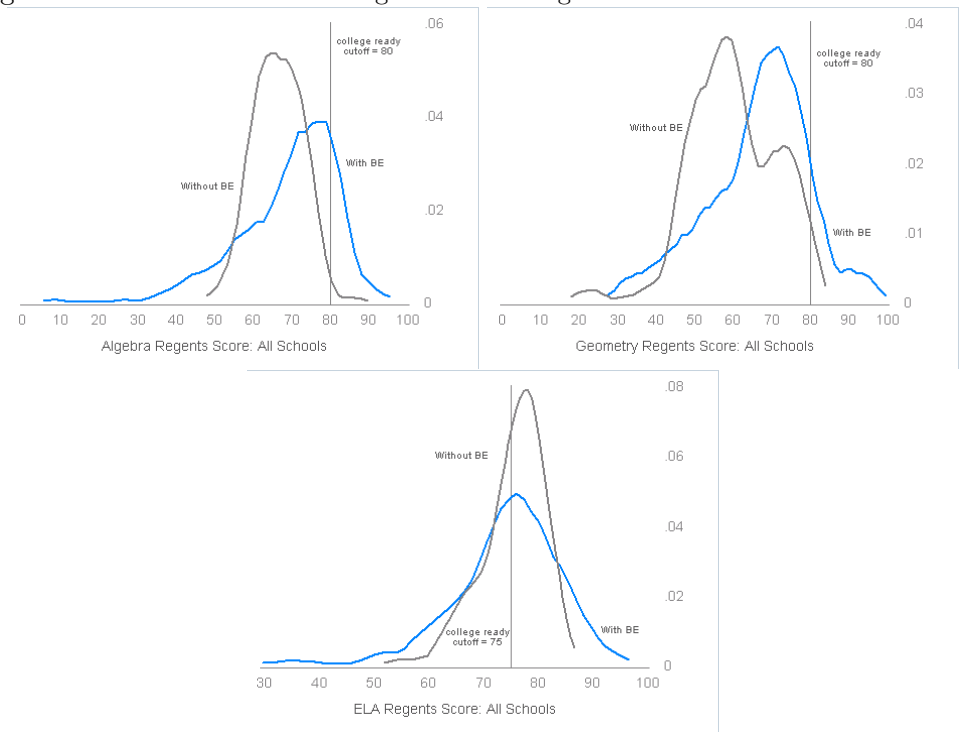
greatly across schools. At School 1, average scores and pass rates were actually significantly lower than predicted, while the portion scoring college ready was significantly greater than predicted. Closer examination of the range and distribution of Geometry scores among School 1 students revealed a much longer tail at the lower-end of the distribution of scores than predicted, meaning that, even though there were more students who scored college ready than predicted, there were also many students who scored much lower than predicted.

Geometry students at School 2 and School 3 scored significantly better than predicted on all three outcomes. Between School 2 and School 3, only 19 (out of 176) students were predicted to pass and only one was predicted to score college ready. In reality, 110 students from these schools passed the exam and 16 scored above the college ready threshold.

Students at School 1 and School 2 performed well on the ELA exam, with 96.1% and 87.0% passing, respectively, though the predicted scores were also very high: that is, students did not perform better than expected. School 1 students scored, on average, almost exactly as predicted (mean score of 76.9 compared to the predicted score of 76.3), while School 2 students scored 2.8 points lower. At both schools, fewer students scored above the college ready threshold than predicted: at School 1, 71% percent scored college ready (78% predicted), while at School 2 44% scored college ready (56% predicted).

Recall that Blue Engine is also interested in understanding its impact on the distribution of scores– i.e., whether students are scoring higher than predicted but also whether students are showing less variability than predicted. Figure 3 shows the distribution of student scores across the three subjects, aggregated by school. These charts suggest that, while the plots of actual scores are further to the right than the predicted scores (with the exception of ELA), the range and variation of scores across all three subjects was larger than predicted. (This was

Figure 3: Distribution of Blue Engine student Regents scores: Predicted versus actual.



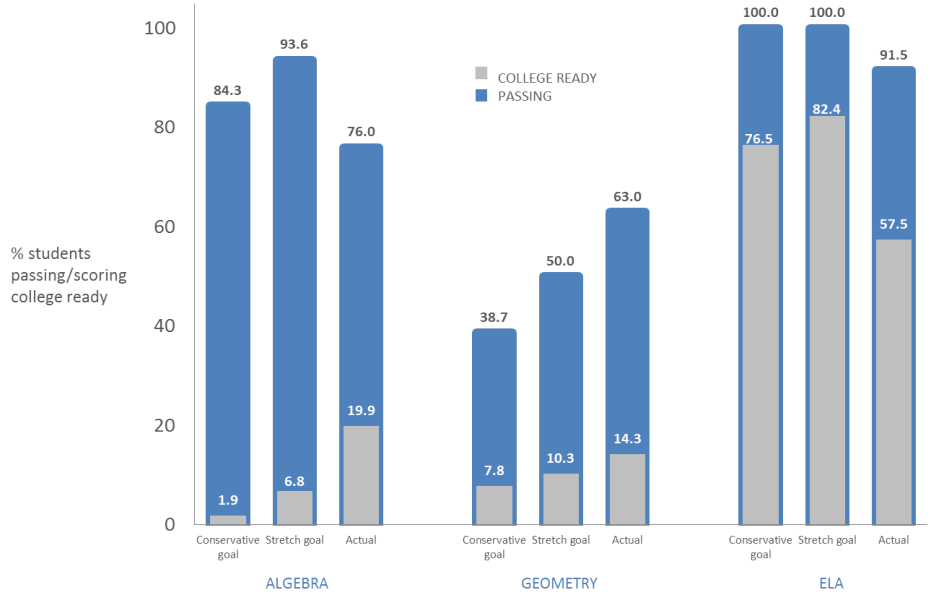
not the case among Geometry students at School 2 and School 3, where students performed much better than predicted, but it was certainly the case at School 1.)

Using predicted scores to set internal goals

The predicted scores can also be used by to set goals for Blue Engine students' performance. The predicted scores offer a baseline forecast of how students should perform on the Regents exams given how similar students have performed in the past. A nonprofit organization that aims to improve students' academic knowledge and skills will want its students to score higher, on average, than these predictions. But how much higher?

One way of determining how much higher students "should" score is to examine how much students' predicted scores typically vary from their actual scores in the model, and then use that information to estimate the likelihood that the difference between a student's predicted score and his or her actual score is more than might be expected due to chance. We create conservative and stretch goals for each student in the Blue Engine program by drawing on something called the standard deviation of the residual. A residual is simply the difference between a student's actual score and his/her predicted score; for example, if a student is predicted to score a 65 and he actually scores a 70, then the residual for this particular student is 5. If he is predicted to score a 65 and he actually scores a 60, then the residual for this student is -5. Like any continuous variable, we can compute summary statistics for these

Figure 4: Blue Engine student performance on Regents exams relative to conservative and stretch goals.



residuals, including the standard deviation, which is a measure of how widely they vary from their mean.

Once we have the standard deviation, we can generate base and stretch goals. It would be unlikely (though clearly not impossible) for a student to score more than one standard deviation higher than predicted; statistically speaking, only 16% of students citywide will do so. A more realistic, though still rigorous, goal would be for students to score .5 standard deviation units higher than predicted. We call this our "stretch" goal. We also set a conservative goal for students to score, on average, .25 standard deviation units higher than predicted. As an example, the standard deviation of the forecast for the 2011-2012 algebra model is 9.5. Thus, a student who is predicted to have a baseline score of 70 on the algebra Regents exam would have a base goal of 72 (70 + .25*9.5) and a stretch goal of 74 (70 + .5*9.5). (For all predictions, we round down to the nearest whole number.)

Table 6 shows the standard deviation of the forecast for each model as well as Blue Engine's conservative and stretch goals for the portion of students passing and scoring college ready in each subject in the 2012-2013 academic year. Figure 4 plots how students actually performed relative to these goals, aggregated across schools. Algebra students did not reach either the base or stretch goals for passing, though nearly three times as many students scored college ready than their stretch goals predicted. In geometry, students exceeded stretch goals for both passing and college ready. Finally, despite high scores among Blue Engine's ELA students, they did not exceed either the conservative or stretch goals.

Table 6: Blue Engine student goals for portion of students passing and scoring college ready.

	% Passing			% College ready				
	Baseline +.0 SD	Conservative +.25 SD	Stretch +.5 SD	Baseline +.0 SD	Conservative +.25 SD	Stretch +.5 SD	N	Forecast SD
Algebra								
School 1	100.0	100.0	100.0	0.0	1.7	6.8	59	10.5
School 2	66.3	79.6	95.9	0.0	0.0	1.0	98	9.6
School 3	70.9	80.0	88.2	1.8	3.6	13.6	110	9.5
All	75.7	84.3	93.6	0.8	1.9	7.5	267	
Geometry								
School 1	90.3	93.6	96.8	6.5	27.4	40.3	62	11.2
School 2	0.0	2.6	6.6	0.0	0.0	0.0	76	11.4
School 3	19.0	32.0	54.0	1.0	2.0	3.0	100	11.4
All	31.5	38.7	50.0	2.1	8.0	11.8	238	
ELA								
School 1	100.0	100.0	100.0	77.6	80.3	84.2	76	11.5
School 2	97.4	100.0	100.0	55.8	72.7	87.0	77	11.8
All	98.7	100.0	100.0	66.7	76.5	85.6	153	

Recommendations Moving Forward

The methods described herein represent a first step in using district-wide data to generate models forecasting Blue Engine students' performance. Below are some recommendations for how the method can be improved going forward.

Individualized student data from a given academic year are typically provided by the New York City Department of Education in the fall following the end of the academic year (e.g., in Fall 2013 for AY 2012-2013). Once these data are available for AY 2012-2013, we can supplement our analysis in a couple of ways. First, we will be able to request several additional years of historical data that we can use to both test our models and to generate school fixed effects based on multiple years of data. For instance, we will be able to determine how well the data from previous years predict performance in 2012-2013. Moreover, as it stands, the fixed effects presented in this paper are based on one year of data and it is plausible that that particular year was an aberration at a given school (i.e., students either scored much higher or lower on the test compared to how they typically score). Using multiple years of data to generate the school fixed effects will give us a better measure of how students in particular schools usually score on exams.

When the data are available, we will also be able to compare differences between Blue Engine students' predicted and actual scores to differences between predicted and actual scores of

students at similar schools. This will allow us to contextualize Blue Engine students' performance. (Note: when the data are available, we will provide scatter plots showing students' actual versus predicted scores, by school.)

The models would also be strengthened by the inclusion of additional predictor variables. Most importantly, adding additional variables for students' previous test scores can improve the models significantly. For instance, the geometry models explain the most variance (i.e., have the highest R-squared values). This is because we used eighth grade scores as well as algebra Regents scores as predictors. Moving forward, we should aim to include as much information as possible for the students. For this project, we were limited by the availability of consistent input data across schools for the Blue Engine students. However, since Blue Engine is currently investing in a database that will act as a reliable repository for student data across schools, we believe we will be able to improve our modeling in the future.

Finally, it is worth warning against potential misuse of these forecasting methods. The student forecasts can be used in a variety of ways among Blue Engine's internal staff, but we warn against sharing the forecasts with teachers and BETAs. There is evidence from social psychological research of a "Pygmalion Effect" for student performance whereby teachers' expectations for how students should perform (given previous information about these students) have predictive power for students' actual performance. In other words, teachers often unknowingly teach students differently depending on their expectations for them. We therefore recommend keeping the forecasts internal.

Appendix A: Non-Fixed Effects Predictions

Table A1 shows predictions from models that did not account for school fixed effects. In schools where the fixed effect was small, the results do not differ much from the results presented in Table 5. In schools where the fixed effect is large (e.g., School 2 geometry), the results differ quite significantly. As a general rule, schools with positive school fixed effects have lower predicted scores and lower portions predicted to pass and score college ready in the non-fixed effects models, while the opposite is true in schools with negative school effects. For instance, the school fixed effect for the algebra models for School 1 students was 2.748. In the non-fixed effects models (where this positive school effect is essentially removed), School 1 students were predicted to perform roughly two points lower and had a lower share of students predicted to pass. Conversely, the school effect for School 2 geometry students was negative and very large, in absolute terms: -17.888. Thus, in the non-fixed effects models, predicted scores as well as the portions predicted to pass and score college ready were much higher than in the fixed effects models.

Table A.1: Comparing predicted versus actual scores among Blue Engine students.

	Mean score		% Passing		% College ready		
	Predicted	Actual	Predicted	Actual	Predicted	Actual	N
Algebra							
School 1	67.2	71.0	89.8	79.7	0.0	23.7	59
School 2	66.3	63.4	67.4	58.2	0.0	9.2	98
School 3	66.9	74.0	62.7	90.9	3.6	27.3	110
All	66.7	69.4	70.4	76.4	1.5	19.9	267
Geometry							
School 1	68.1	68.0	69.4	64.5	0.0	29.0	62
School 2	68.1	62.7	71.1	48.7	5.3	5.3	76
School 3	60.9	68.2	28.0	73.0	3.0	12.0	100
All	65.1	66.4	52.5	63.0	2.9	14.3	238
ELA							
School 1	73.4	76.9	98.7	96.1	42.1	71.1	76
School 2	71.7	71.6	94.8	87.0	29.9	44.2	77
All	72.6	74.3	96.7	91.5	36.0	57.5	153

Appendix B: Missing data

Table B.1: Accounting of missing data among 2012-2013 Blue Engine students.

	Total # of students	# absent	# missing predictions	Total # missing	# used in analysis
Algebra					
School 1	61	1	1	2	59
School 2	103	4	1	5	98
School 3	115	2	3	5	110
Total	279	7	5	12	267
Geometry					
School 1	73	11	0	11	62
School 2	118	41	1	42	76
School 3	106	6	0	6	100
Total	297	58	1	59	238
ELA					
School 1	79	3	0	3	76
School 2	81	2	2	4	77
Total	160	5	2	7	153