Estimating the Impact of the Blue Engine Program on Common Core Algebra Scores

February 2015

Rebecca Casciano

Glass Frog Solutions

rebecca@glassfrog.us

Abstract

Blue Engine is an education organization that places teaching assistants in classrooms at New York City partner high schools to provide instructional support and tutoring, facilitate extra learning time for students, and implement a social cognitive curriculum. This report uses quasi-experimental methods to estimate the effect of the Blue Engine intervention on Common Core algebra test scores in 2013-2014. Specifically, we use a propensity score estimation method to match Blue Engine ninth graders to a sample of similar students attending similar schools. We supplement the analysis by estimating the impact of the intervention using multi-level regression methods. The results indicate that average scores and pass rates were significantly higher among Blue Engine students than among the matched sample. Blue Engine students' average scores were roughly .39 to .44 standard deviation units higher than their peers, depending on the methodology used. We observed a significant impact on college ready rates using the matching methodology, but not the multi-level regression approach. The results suggest that Blue Engine may be a promising intervention for helping students learn Common Core algebra content knowledge and prepare for more rigorous tests. By providing students with more individualized instruction and extra learning time, Blue Engine may be able to close the gap for students who enter high school at a deficit relative to their peers.

Introduction

This report presents results from a project conducted on behalf of Blue Engine, a nonprofit 501(c)(3) charitable organization that places teaching assistants (called BETAs) in classrooms at New York City partner high schools. BETAs work alongside math and English teachers within these schools to provide tutoring and extra support, facilitate extra learning time for students, and implement a curriculum designed to improve students' social cognitive learning. The overarching goal of the program is to help students be admitted to, enroll in, and graduate on time from college.

In academic year (AY) 2013-2014, Blue Engine was in its fourth year of service, working with 740 students in algebra, geometry, and ELA across five schools. The program measures its impact annually by assessing student scores on New York State Regents exams in these subjects. One challenged it faced in 2013-2014 was implementing its program and assessing its impact amidst the state's transition to Common Core (CC) standards and assessments. AY 2013-2014 was the first year that a large share of schools opted to sit students for the CC algebra exam. Three of Blue Engine's five partner schools chose to sit all or some of their students for this exam. Across both the city and state, scores and pass rates were lower on the CC algebra exam than they have historically been on the Integrated Algebra (IA) exam. For programs like Blue Engine, it is difficult to contextualize their impact due to a lack of historical benchmarking data.

Blue Engine commissioned Glass Frog Solutions to estimate how Blue Engine students might have performed on the CC algebra exam in the absence of their work with Blue Engine. We draw on student- and school-level data from the New York City Department of Education (NYC DOE) to match Blue Engine students to similar students attending similar schools but not working with the program; we then compare performance on the Regents exam between Blue Engine students and this matched comparison sample. We supplement this analysis with a multi-level regression analysis that estimates the impact of the Blue Engine treatment on student outcomes, controlling for student- and school-level characteristics, and accounting for the nested structure of the data.

We begin by describing the study sample, measures, and methodology. We then describe the results and offer a brief discussion in light of public discourse on the transition to Common Core testing.[1]

[1] This report is intended for an internal audience that is familiar with Blue Engine's mission and program model and has some familiarity with the organization's model and practices around evaluation and performance measurement. Questions and additional requests for information should be directed to the study's author.

Study Sample

In AY 2013-2014, Blue Engine worked with 199 students across three partner schools in CC algebra classrooms.[2] All of these students are included in this analysis, with the exception of two groups.

First, we exclude 18 students who were in tenth to twelfth grade or were repeating the ninth grade. Blue Engine works primarily with ninth grade students taking the exam for the first time. They sometimes work with students in other grades who are retaking the exam. We exclude non-ninth graders from this analysis since it is likely they had previously taken algebra, possibly in a Blue Engine classroom, and most likely preparing for the IA exam. This makes it difficult to isolate the specific impact of Blue Engine's *Common Core* support on student performance.

Second, Blue Engine internally considers a student part of its treatment population if s/he is in a Blue Engine classroom (that is, a classroom to which BETAs are assigned) and if his/her attendance is greater than 50 percent. Students below this 50 percent threshold are excluded from all analyses. To be consistent with how Blue Engine performs its internal analyses, we therefore exclude two students with attendance below this threshold.

For comparative purposes, we also exclude non-ninth graders and students with attendance rates below 50 percent from the comparison population.

The final Blue Engine sample includes 179 students. These results can be generalized to ninth grade students with attendance above 50 percent.

Table 1 shows the number of students included in the analysis across the three schools, as well as their background characteristics; for comparison, we also show characteristics of students attending Peer Horizon schools and other district schools.[3] On average, Blue Engine students are comparable to students at Peer Horizon schools, though they have slightly lower eighth grade test scores, are more likely to be Hispanic, and are less likely to be black. Relative to students attending other district schools, Blue Engine students have much lower eighth grade scores, are more likely to have IEPs, and are more likely to be Hispanic. These estimates indicate that Blue Engine is serving students who are starting high school with significant learning needs and/or who are already behind their peers districtwide.

[2] As mentioned, New York State allowed schools to decide which Algebra exam students would prepare for. Two Blue Engine schools prepared exclusively for the IA exam, two prepared exclusively for the CC exam, and one school took a mixed approach, having some students take the IA exam and others take the CC exam.

[3] For each public school in New York City, the DOE establishes a set of roughly 40 comparable schools based on the characteristics of incoming students; together, these 40 schools comprise a school's "Peer Horizon." In the impact analysis, we limit our comparison sample to students attending Peer Horizon schools, which is why we present data from those schools separately here.

Table 1: Average characteristics for ninth grade students taking Common Core Algebra exam in AY 2013-2014. (PH = Peer Horizon) Source: 2013-2014 Student Biographical Data, New York City Department of Education.

	Blue Engine schools					
	School 1	School 2	School 3	All	PH schools	Other district schools
Ave. grade 8 ELA score	285.0	278.0	273.7	277.2	283.0	292.4
Ave. grade 8 math score	295.2	280.3	272.8	279.3	286.0	298.2
% ELL	22.2	1.3	14.3	10.1	11.0	12.5
% IEP	7.4	24.0	22.1	20.7	19.0	13.6
% black	7.4	28.0	31.2	26.3	35.5	29.0
% Hispanic	92.6	61.3	61.0	65.9	51.3	36.8
% free/reduced lunch	77.8	74.7	85.7	79.9	82.2	76.1
% female	25.9	58.7	53.2	51.4	46.2	50.2
N	27	75	77	179	10,069	26,057

Using data from the NYC DOE, we estimate Blue Engine's impact on three student outcomes:

- *Regents score*: The student's scaled score on the 2013-2014 CC Algebra Regents exam. Scores range from 0 to 100. If a student took the exam more than once during the year, we used the student's highest score. In AY 2013-2014, the average unadjusted score among students attending peer schools was 58.4; among Blue Engine students, the average score was 62.0.
- *Pass rate*: The second outcome we consider is the proportion of students passing the CC Algebra exam. Students pass Regents exams with a minimum score of 65; special education students pass with a score of 55. For each school, the proportion of students passing is equal to the number who scored at or above the passing threshold divided by the total number of students who took the exam. The average pass rate among students attending peer schools was 39.8 percent; the average pass rate among Blue Engine students was 56.4 percent.
- *College ready rate*: The third outcome we consider is the proportion of students scoring above the college ready threshold on the CC Algebra exam. The city uses a threshold of 70 for all students; there is not a separate threshold for special education students. The proportion of students scoring college ready is equal to the number who scored 70 or higher divided by the total number of students who took the exam. At peer schools, 15.4 percent of students scored above the college ready threshold in AY 2013-2014, compared to 15.1 percent of Blue Engine students.

Methodology

As shown in Table 1, Blue Engine students differ in some ways from students at Peer Horizon schools. This can be problematic if these characteristics are also associated with scores on Regents exams, making it difficult to determine whether participating in Blue Engine or some other factor is the reason for differences on the outcomes of interest. To control for underlying differences between Blue Engine and non-Blue Engine students, we employed propensity score estimation procedures to establish credible comparison groups. Propensity score methods attempt to model the treatment assignment process in an effort to identify individuals who are "similar" to each other on variables that influence both treatment assignment and outcomes.[4] Specifically, we use the propensity scores to identify members of the comparison group (i.e., students attending Peer Horizon schools) who have similar propensity scores to members of the treatment group (i.e., Blue Engine students). We then compare scores only between students who have similar propensity scores.

The matching procedure involved the following steps:

1. As described above, we began by limiting the matching population[5] to ninth grade students who attended a Peer Horizon school in AY 2013-2014 and whose attendance rate was greater than 50 percent.
2. We used probit regression models to estimate each student's individual likelihood of being in the Blue Engine program. Probit regression models are used when the outcome you are estimating is binary (yes/no). In this case, students either participated in the program (yes) or they did not participate (no), so a probit model is appropriate. The probit model yields, for each student, the probability (range: 0 – 1.0) that they participated in Blue Engine. Output from these models is available upon request. The regression models included the following predictor variables:
    - Whether the student has an IEP (yes/no)
    - Whether the student is an English language learner (ELL) (yes/no)
    - Whether the student is female (yes/no)
    - Whether the student is eligible for free/reduced price lunch (yes/no)
    - Student ethnicity: Students could be identified as Asian, Black, Hispanic, Multi-Racial, Native American, or White. We included ethnicity as a series of dummy variables indicating which ethnic group the student belonged to.

[4] Rosenbaum, P.R. & Rubin, D.B. (1983). "The central role of the propensity score in observational studies for causal effects." Biometrika, 70(1): 41-55.

[5] We call the population from which we sampled students for the comparison group the "matching population."

- Student ELA and math eighth grade state exam scores, as well as squared-terms on these measures
- Whether eighth grade math and English scores were imputed: Across all Blue Engine and Peer Horizon schools, 7.5 percent of students were missing eighth grade math scores and 8.4 percent were missing ELA scores. For these students, we imputed scores to be equal to the mean of students in their cohort at their current school. We created binary variables indicating whether students were missing scores in ELA and math and included both indicators in the model.
- A series of interaction variables: ELA8 missing * ELL status; MATH8 missing * ELL status; ELA8 missing * school-level ELA proficiency level; MATH8 missing * school-level math proficiency level; IEP status * MATH8; ELL status * MATH8; ELL status * ELA8; free/reduced price lunch * MATH8

3. Once we estimated each student's likelihood of being in Blue Engine, we matched students to non-Blue Engine students in the matching population on the propensity scores, using an approach called nearest-neighbor matching. Specifically, we matched each Blue Engine student to a non-Blue Engine student whose propensity score was within +/- 0.01 points of their own score. (As an example, a student with a propensity score of .13 would be matched to another student with a score ranging from 0.12 – 0.14.) We matched with replacement since the distribution of propensity scores differed between groups, with non-treatment students having fewer cases at the upper-end of the score distribution.[6] One Blue Engine student had a propensity score outside the region of common support and was dropped from the analysis. This method yielded a final sample of 160 non-Blue Engine students, weighted such that each of the 178 Blue Engine students in the sample has one (not necessarily unique) match. The mean and variance of the propensity scores are nearly identical between groups $[\mu_{BE} = .092, \sigma_{BE} = .072; \mu_{PH} = .092, \sigma_{PH} = .072]$.

4. The goal of this process was to create two samples that were statistically similar (or "balanced") on the variables we would expect to impact student outcomes. To determine whether the two samples were balanced on these variables, we tested for mean differences between the Blue Engine students and non-Blue Engine students on each variable. Results from the t-tests for the final, balanced samples are reported in Table 2. As these tables show, across all groups, the matching created balanced samples that do not differ statistically or substantively.

5. When we were certain that the two samples were comparable, we compared them on the outcome measures described in the previous section using ordinary least squares regressions and applying the propensity score frequency weights.[7]

[6] See Dehejia, Rajeev, Wahba, Sadek, 2002. Propensity score matching for nonexperimental causal studies. The Review of Economics and Statistics 84(1), 151-161.

[7] The students are clustered within schools and therefore the observations are not independent; we executed two-level random intercept mdoels, but the proportion of variance explained at the school-level was both statistically and

Table 2: Mean scores on key covariates for Blue Engine students and matched sample, along with results from between-groups t-test. Source: 2013-2014 NYC Department of Education student-level data.

	Blue Engine	Matched sample	t-score	p-value
Mean propensity score	0.1	0.1	-0.004	0.996
Student-level characteristics				
Ave. grade 8 math score	279.3	280.9	-0.644	0.520
Ave. grade 8 ELA score	277.2	279.1	-0.714	0.476
% IEP	20.8	18.0	0.669	0.504
% ELL	10.1	12.4	-0.670	0.503
% black	26.4	26.4	0.000	1.000
% Hispanic	65.7	66.3	-0.112	0.911
% free/reduced lunch	80.3	79.2	0.263	0.793
% female	51.1	47.2	0.741	0.459
% ELA8 missing	11.2	11.2	0.000	1.000
% MATH8 missing	9.0	9.0	0.000	1.000
School-level characteristics				
Ave. ELA proficiency level	2.3	2.3	-0.967	0.334
Ave. math proficiency level	2.2	2.2	-1.762	0.079
% students with disabilities	22.7	22.5	0.655	0.513
% black or Hispanic	90.4	91.4	-1.555	0.121
% ELL	10.0	9.7	0.417	0.677
Ave. teacher absences (days)	5.0	5.0	0.211	0.833
Ave. years teaching experience	6.2	6.3	-0.681	0.496
N	178	178		

substantively insignificant, so we opted to use standard OLS models. The reason we use OLS regressions instead of t-tests to compare means between groups is one of practicality: we could not apply the frequency weights to the means using Stata's ttest commands.

Results

Outcome measures, as well as 95 percent confidence intervals, are reported in Table 3. For ease of interpretation, we also present linear predictions graphically for each outcome measure in Figure 1.

Table 3: Results from OLS regressions comparing weighted means among Blue Engine students and matched sample on Common Core Algebra exam in AY 2013-2014. (N = 356) Source: 2013-2014 Student Biographical Data, New York City Department of Education.

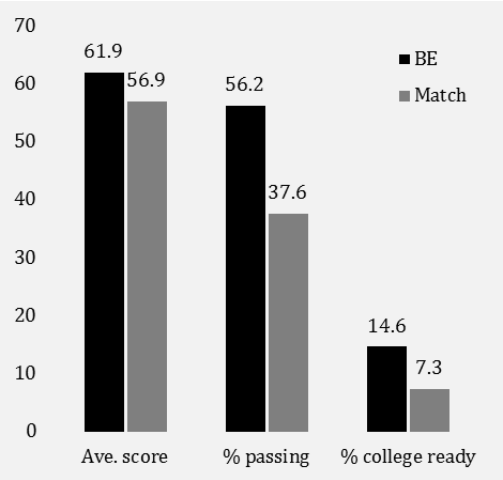
	Average score		Prop. passing		Prop. college ready	
	Coef.	95% CI	Coef.	95% CI	Coef.	95% CI
Blue Engine treatment	4.97	2.846, 7.087	0.185	0.083, 0.288	0.073	0.008, 0.138
Constant	56.938	55.439, 58.438	0.376	0.304, 0.449	0.073	0.027, 0.119

Blue Engine students outperformed students in the matched comparison sample on all three outcomes. Average scores were higher by a margin of five points, the equivalent of roughly .44 standard deviations (the pooled standard deviation for the full sample of Blue Engine and Peer Horizon students is 11.3). Pass rates were higher among Blue Engine students by a margin of 18.5 percentage points; this is roughly equivalent to 33 additional students passing the Common Core exam than would have passed in the absence of Blue Engine. College ready rates were also modestly higher among Blue Engine students: 14.6 percent passed compared to 7.3 percent in the matched sample. This is equivalent to 13 additional students scoring college ready than would have in the absence of Blue Engine.

Alternative methodology

As a robustness test, we also performed multilevel regression models estimating student performance on the state tests. We estimated separate models for each of the three outcomes, using ordinary least squares regression models, controlling for treatment status as well as relevant individual- and school-level characteristics, and including school-level random intercepts to account for clustering within schools. Student-level controls included: eighth grade state math and ELA scores (including squared terms on each measure); IEP and ELL status; whether student is eligible for free/reduced price lunch; whether the student is female; race/ethnicity; and whether the student was missing eighth grade math or ELA scores. We included a binary measure indicating treatment status (i.e., whether the student was in Blue Engine). Finally, we also included a set of school-level controls, including average number of teacher absences, average teacher experience (in years), average math and ELA proficiency, and the percent of students in the following categories: English language

Figure 1: Linear predictions of Regents outcomes for Blue Engine students and matched sample, based on estimates from Table 3. Source: 2013-2014 Regents Data, New York City Department of Education.



learners, special education, self-contained, overage, and black or Hispanic. The final equation for estimating the adjusted mean difference between Blue Engine outcomes and outcomes among Peer Horizon students is given as follows for student i in school j:

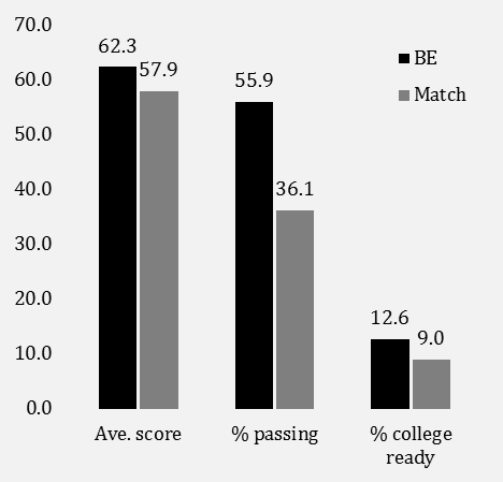
$$y_{ij} = \beta_1 + \beta_2 x_{2ij} + \delta \mathbf{x}_{ij} + \gamma \mathbf{x}_{ij} + \zeta_j + \epsilon_{ij}$$

where x_{2ij} is a binary variable indicating treatment status (Blue Engine = 1), $\mathbf{x}_{ij}$ is a vector of student-level covariates, $\mathbf{x}_{ij}$ is a vector of school-level covariates, ζ_j is the residual shared by students attending the same school, and ϵ_{ij} is the residual unique to each student. This model allows us to estimate the difference between Blue Engine students' scores and non-Blue Engine students' scores, controlling for all of the variables in the model. The results are presented in Table 4. Once again, to assist with interpretation of the results, we present linear predictions in Figure 2.

The results are similar to the results from the propensity score matching estimation, with the exception of college ready rates. Holding the other variables in the model at their means, Blue Engine students scored 4.4 points higher on the exam than students attending Peer Horizon schools ($\mu_{BE} = 62.3, \mu_{PH} = 57.9$). This translates into a standardized gain of .39 standard deviation units. The pass rate was 19.8 percentage points higher among Blue Engine students than among comparison students ($\mu_{BE} = 55.9\%, \mu_{PH} = 36.1\%$), indicating that 35 additional students passed the exam than would have in the absence of the program. Finally, using this modeling approach, college ready rates were only 3.7 points higher among Blue Engine students than among students

attending peer schools ($\mu_{BE} = 12.6\%, \mu_{PH} = 9.0\%$). This translates into 6.6 additional students scoring college ready, though this difference was not statistically significant.

Figure 2: Linear predictions of Regents outcomes for Blue Engine students and Peer Horizon students, based on estimates from Table 4. Source: 2013-2014 Regents Data, New York City Department of Education.



Summary and Discussion

To summarize, using both the propensity score methodology and a multilevel regression approach, we observed a substantively and statistically significant impact of the Blue Engine treatment on student scores and pass rates on the CC algebra exam. Depending on the methodology, Blue Engine students' scores were .39 to .44 standard deviation units higher than their peers, the equivalent of roughly 4.0 to 5.0 points on a 100 point scale. Pass rates were also higher by a margin of roughly 18 to 20 percentage points. Using the propensity score matching approach, we observed a significant difference in college ready rates between Blue Engine students and the matched sample; the multilevel regression approach yielded a smaller and statistically insignificant effect size.

The propensity score methodology arguably offers a more valid estimate of Blue Engine's impact, given that it compares Blue Engine students only to other students with similar characteristics attending schools that are similar on observable characteristics. Thus, it is encouraging that this methodology yielded positive results on the college ready outcome. It is unclear why Blue Engine students were not able to see larger gains on the college ready outcome. Among all ninth grade New York City students who took the exam, the average score was 61.2, the pass rate was 50.7 percent, while the college ready rate was 26.0 percent. This means that Blue Engine students were able to improve their scores enough so that even their unadjusted average pass rate exceeded that of the district overall, yet they still lagged behind the district in college ready rates. Additional research

is needed to determine why Blue Engine students were unable to see larger gains on this outcome, as well as how Blue Engine might alter its design to improve college ready rates going forward.

At the time of this report, both local and national education organizations debated the merits of the Common Core curriculum and the attempts to assess Common Core standards using traditional standardized tests. Relative to historical performance, the lower scores and pass rates on the CC test are evidence that districts and schools are still figuring out how to prepare students for these more rigorous tests. For schools serving students with disproportionately economically disadvantaged or special education populations, this undoubtedly presents an even greater challenge.

The results presented in this report suggest that Blue Engine may be a promising intervention for helping students learn CC algebra content knowledge and prepare for more rigorous tests. As shown in Table 1, Blue Engine students have lower eighth grade scores, are more likely to have IEPs, and are more likely to be Hispanic than students districtwide, yet their scores and pass rates were still on par with students districtwide. By providing students with more individualized instruction and extra learning time, Blue Engine may be able to close the gap for students who enter high school at a deficit relative to their peers districtwide.

Table 4: Adjusted Regents outcomes for Blue Engine students, using multilevel regression methods. Source: 2013-2014 NYC Department of Education student-level data.

	Average score			Prop. passing			Prop. college ready		
		95% CI			95% CI			95% CI	
	Coef.	Lower	Upper	Coef.	Lower	Upper	Coef.	Lower	Upper
Blue Engine treatment	4.424	0.785	8.063	0.198	0.047	0.348	0.037	-0.093	0.166
8th grade math score	-0.287	-0.366	-0.208	-0.017	-0.021	-0.013	-0.036	-0.039	-0.033
8th grade math score (squared)	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000
8th grade math missing	-1.061	-2.656	0.535	-0.039	-0.115	0.037	0.036	-0.019	0.091
8th grade ELA score	0.043	-0.029	0.116	-0.007	-0.010	-0.003	-0.008	-0.011	-0.006
8th grade ELA score (squared)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
8th grade ELA missing	0.761	-0.773	2.295	0.045	-0.028	0.118	0.040	-0.012	0.093
IEP	-2.937	-3.402	-2.472	0.257	0.235	0.280	-0.012	-0.028	0.004
ELL	-1.901	-2.493	-1.309	-0.030	-0.058	-0.002	-0.018	-0.038	0.002
Female	0.484	0.135	0.833	0.011	-0.006	0.028	0.004	-0.008	0.016
Eligible free lunch	-0.010	-0.453	0.432	0.008	-0.013	0.029	-0.016	-0.031	0.000
Ethnicity (ref = White or Other)									
Asian	0.834	-0.103	1.771	0.061	0.017	0.106	0.044	0.012	0.077
Black	-0.953	-1.678	-0.229	-0.038	-0.072	-0.003	-0.041	-0.066	-0.016
Hispanic	-0.695	-1.398	0.009	-0.036	-0.069	-0.002	-0.041	-0.065	-0.017
School-level variables									
Ave. ELA proficiency level	4.860	-6.625	16.346	0.158	-0.314	0.631	0.151	-0.258	0.559
Ave. math proficiency level	-1.018	-10.084	8.047	0.013	-0.361	0.386	0.107	-0.216	0.429
% students with disabilities	-0.024	-0.227	0.179	-0.003	-0.011	0.005	-0.003	-0.011	0.004
% self-contained	0.102	-0.140	0.343	0.004	-0.006	0.014	0.004	-0.004	0.013
% overage	-0.367	-0.682	-0.051	-0.012	-0.025	0.001	-0.003	-0.014	0.008
% black or Hispanic	0.009	-0.056	0.075	0.000	-0.002	0.003	0.001	-0.001	0.003
% ELL	0.036	-0.065	0.137	0.002	-0.003	0.006	0.001	-0.002	0.005
Ave. teacher absences (days)	-0.088	-0.386	0.211	-0.006	-0.018	0.007	-0.005	-0.016	0.006
Ave. years teaching experience	-0.019	-0.235	0.196	0.002	-0.007	0.010	0.000	-0.008	0.008
Constant	51.320	21.458	81.183	1.907	0.637	3.176	4.869	3.814	5.924
N	10,248			10,248			10,248		
Derived estimates									
R^2	0.35			0.24			0.25		
ρ	0.12			0.09			0.13		