Firecracker Step 1 Performance Analysis

Elliott Bartsch
Data Scientist, Firecracker
January 28, 2015

Introduction

To collect data to assess Firecracker's efficacy as a review source for the USMLE Step 1 exam, we distributed a survey and requested score reports medical students across the US using our database of over 42,000 unique email addresses[1]. We received 2648 completed survey responses, 508 official score reports, and we were able to connect the accounts the scores of 443 students to their Firecracker accounts. To incentivize responses, we gave survey respondents a free month of access to Firecracker's program if they completely filled the survey.

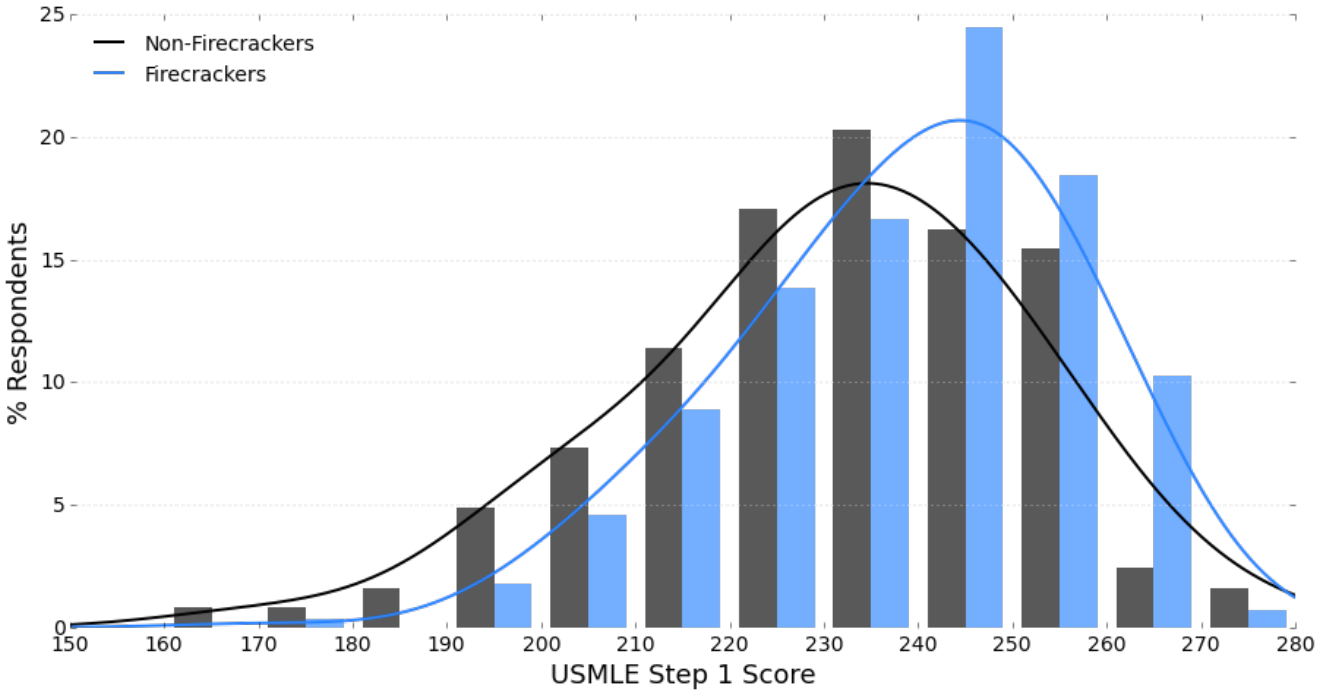
Analysis

We used the following main metrics to gauge Firecracker usage:

Metric	Description
Active Weeks	The number of weeks the given student did at least 100 questions on Firecracker
Recalls Done	The total number of question reviews the student did on Firecracker
Unique Questions	The number of unique questions the student saw on Firecracker
Final Average	The average final score of all the questions the user did

We compared the distributions of non-Firecracker Step 1 scores with the distribution of Firecracker Step 1 scores. We filtered for Firecracker users who actively used the program for at least 12 weeks (*n Firecrackers = 282*). Non-Firecracker scores were found from students who did not indicate using Firecracker in our survey and whose emails were not associated with accounts in our database (*n non-Firecrackers = 123*).

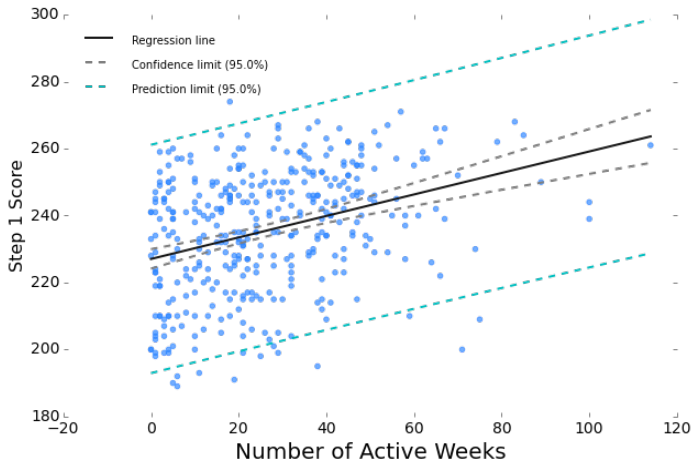
[1] https://firecrackerinc.wufoo.com/forms/firecrackers-annual-medical-student-survey/



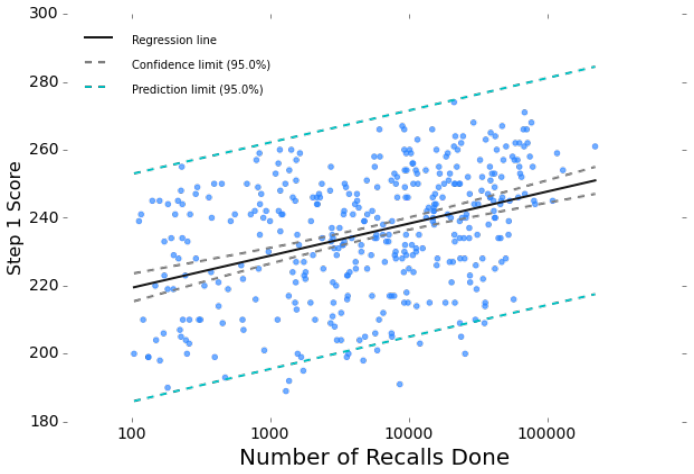
The mean non-Firecracker Step 1 score was a 230.7 with a standard deviation of 20.8. This is right in line with the reported 2014 national statistics[2] (national average of 230 and standard deviation of 20). The mean Firecracker score was a 238.5 with a standard deviation 17.5. As we can see, the distribution of Firecracker scores is shifted right and has a narrower spread. Using an unpooled two sided t-test, we found that the average Firecracker score is significantly higher than the average non-Firecracker score ($p=0.0003$).

We examined how individual usage metrics correlated with performance on the Step 1 exam. Because of the high variance in scores with only a rudimentary use of Firecracker's program, we filtered for users who did at least 100 questions on Firecracker.

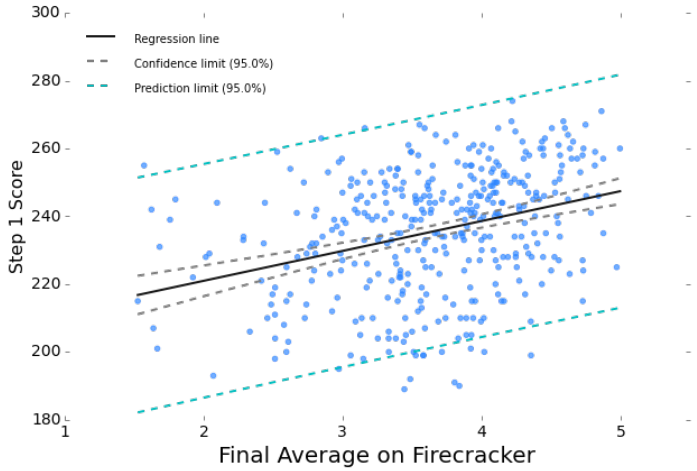
[2] http://en.wikipedia.org/wiki/USMLE_Step_1



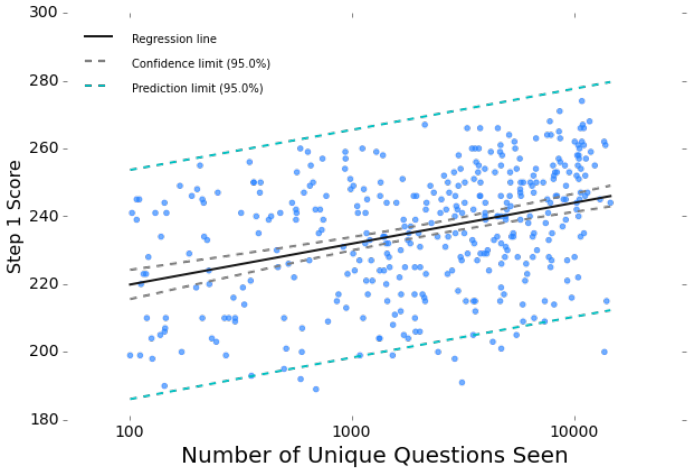
r=0.343, *n*=386, *p*<0.001



r=0.389, *n*=386, *p*<0.001



r=0.330, *n*=386, *p*<0.001



r=0.374, *n*=386, *p*<0.001

We can see that there are significant correlations between each of our usage metrics and USMLE Step 1 scores.

In our survey we also found a significant correlation between MCAT and Step 1 scores (*r=0.414*). This level of correlation has also been replicated in published academic studies[3]. Even after adjusting for MCAT scores, the above Firecracker usage metrics are still significant

[3] Donnon, Tyrone, Elizabeth Oddone Paolucci, and Claudio Violato. "The predictive validity of the MCAT for medical school performance and medical board licensing examinations: a meta-analysis of the published research." Academic Medicine 82.1 (2007): 100-106.

predictors of Step 1 scores. In the following multiple regression we filtered for users who did at least 100 questions on Firecracker's program (*n=386*):

OLS Regression Results

Dep. Variable:	step_1_score	R-squared:	0.346
		Adj. R-squared:	0.339
Method:	Least Squares	F-statistic:	50.48
No. Observations:	386	Prob (F-statistic):	4.26e-34
Df Residuals:	381	Log-Likelihood:	-1598.8
Df Model:	4	BIC:	3227.

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
const	138.4797	7.926	17.472	0.000	122.896	154.063
active_weeks	**0.0950**	0.064	1.483	0.139	-0.031	0.221
log10_recalls_count	**5.5328**	1.701	3.252	0.001	2.187	8.878
final_avg	**4.1337**	1.236	3.344	0.001	1.703	6.564
mcat	1.8456	0.196	9.403	0.000	1.460	2.232

Omnibus:	8.087	Durbin-Watson:	1.831
Prob(Omnibus):	0.018	Jarque-Bera (JB):	8.352
Skew:	-0.355	Prob(JB):	0.0154
Kurtosis:	2.885	Cond. No.	455.

Adding Firecracker usage statistics to the MCAT only model explains an additional 17.5% of the variance in Step 1 scores (R^2=*0.171* for MCAT only model, see appendix). The coefficients for every Firecracker usage statistic are positive. We see that the coefficient for $\log_{10}$ number of recalls is 5.53. Our model estimates that a student who does 1000 recall questions on Firecracker will score 5.53 points higher than student who only does 100 recall questions, even after adjusting for MCAT scores. The more questions students do on Firecracker, the higher their predicted score.

We also see that the difference between having a final average of 4 compared to a 3 on Firecracker's questions results in a 4.13 point difference in predicted Step 1 scores, even after adjusting for MCAT scores. Students that achieve mastery of Firecracker's content also do well on their Step 1 exams. Although not statistically significant at the 0.05 level (*p=0.139*), the coefficient for the number of active weeks is also positive even after adjusting for MCAT scores and the other Firecracker usage metrics. We can see the importance of spacing out review through Firecracker's system instead of cramming.

Appendix

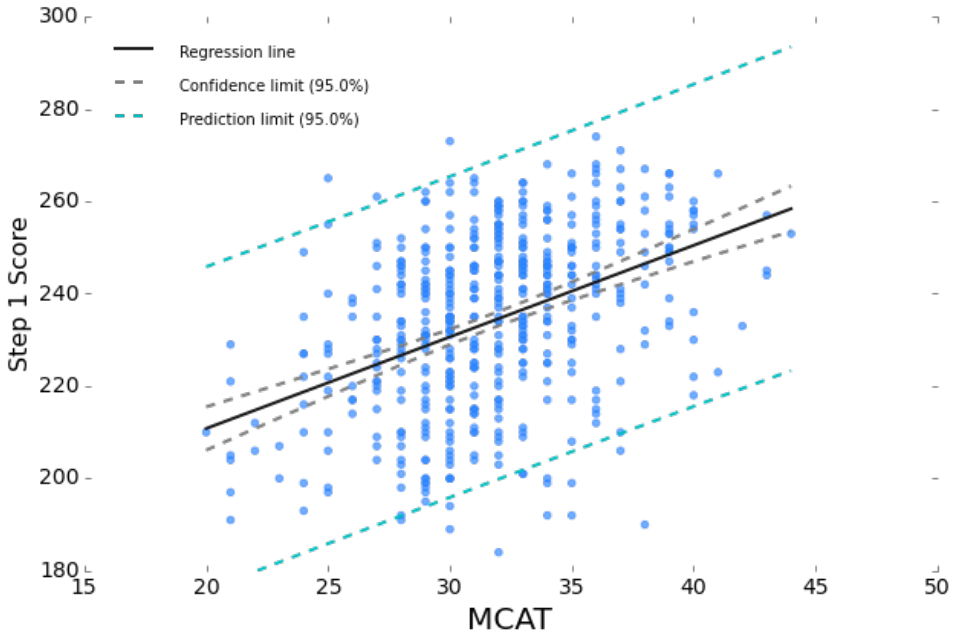
MCAT only model. ANOVA test for model comparison shows that the model with Firecracker usage statistics included significantly explains more of the variance in Step 1 scores ($p<0.001$).

OLS Regression Results

Dep. Variable:	step_1_score	R-squared:	0.171
		Adj. R-squared:	0.169
Method:	Least Squares	F-statistic:	104.8
No. Observations:	510	Prob (F-statistic):	1.73e-22
Df Residuals:	508	Log-Likelihood:	-2202.1
Df Model:	1	BIC:	4417.

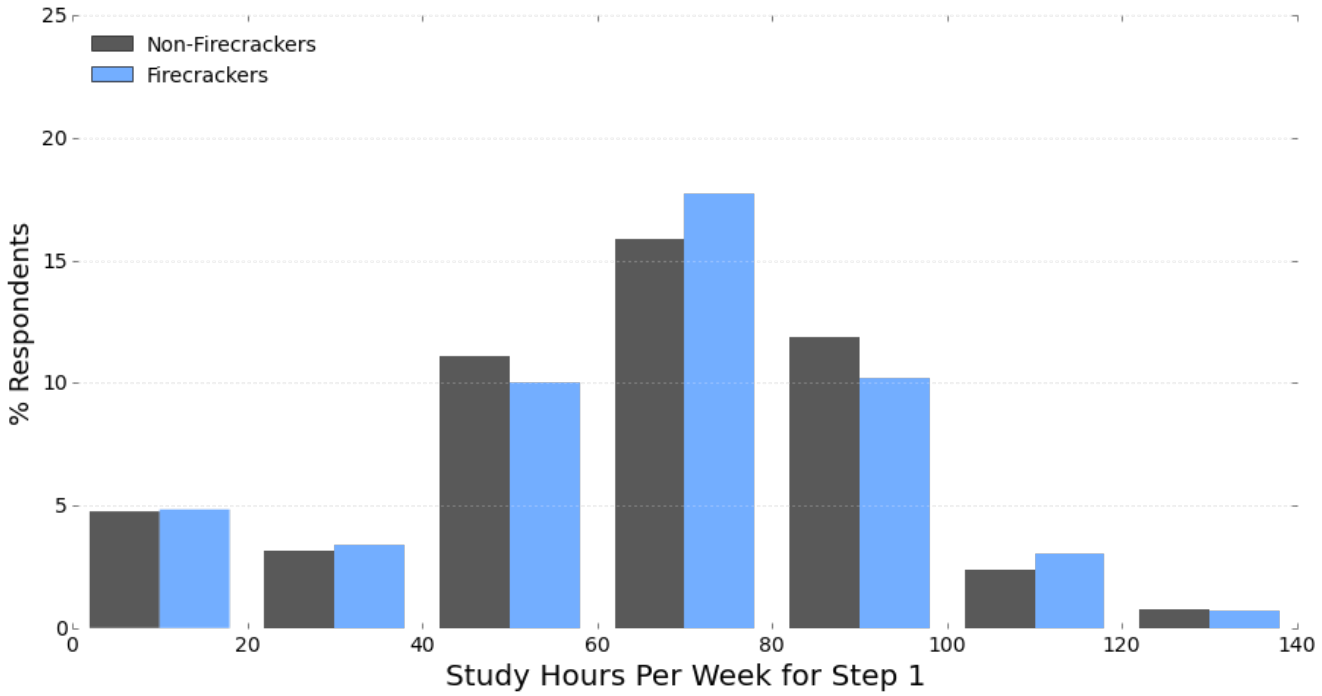
	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
const	169.7566	6.300	26.944	0.000	157.379	182.135
mcat	2.0179	0.197	10.237	0.000	1.631	2.405

Omnibus:	35.701	Durbin-Watson:	1.977
Prob(Omnibus):	0.000	Jarque-Bera (JB):	42.257
Skew:	-0.632	Prob(JB):	6.67e-10
Kurtosis:	3.626	Cond. No.	250.



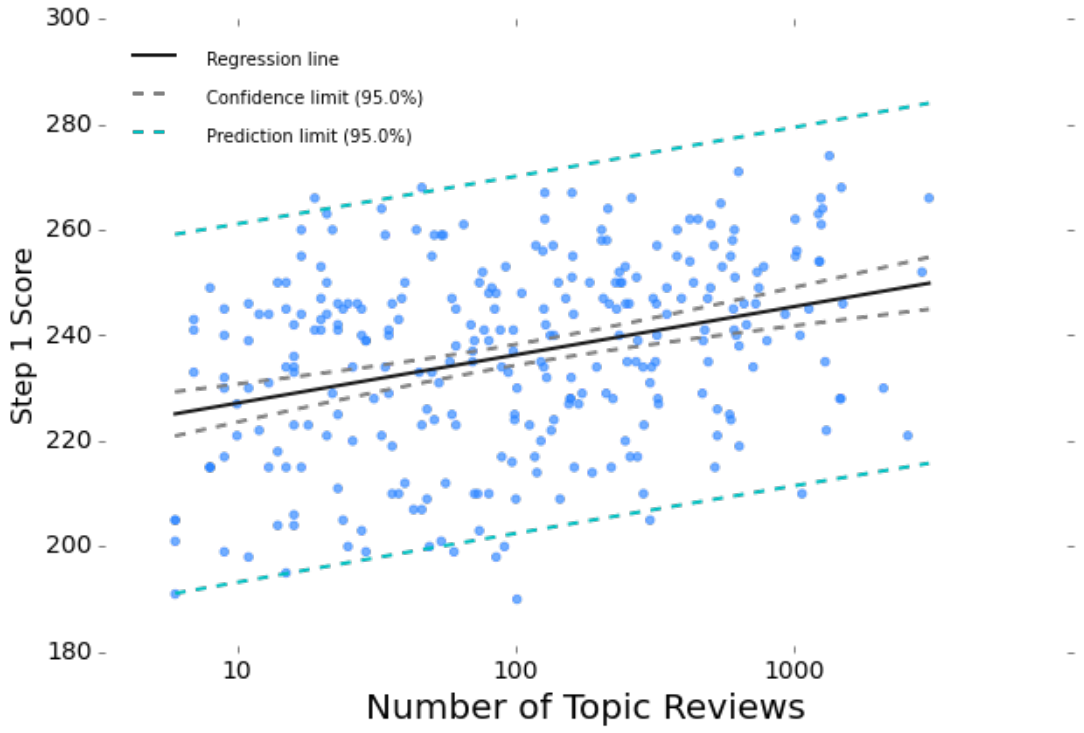
r=0.416, n=508, p<0.001

We found that Firecracker users did not report studying for longer periods of time than non-Firecrackers.



Using a two-sided, pooled t test, we found no significant difference in study times for Step 1 between Firecrackers and non-Firecrackers (*p=0.967*).

We also found a significant correlation between the number of topics reviewed on Firecracker and Step 1 performance. Our system does not track every topic review, so we did not have complete topic review data for every user (n=292):



r=0.330, n=292, p<0.001