Firecracker Step 1 Performance Analysis

Elliott Bartsch
Director of Data Science, Firecracker
September 15, 2016

Introduction

Data is essential for testing and validating Firecracker as an online medical education platform. Firecracker has continuously surveyed medical students for their board exam scores since August 2014. We published our initial analysis of Firecracker's efficacy as a study resource for the USMLE Step 1 and COMLEX Level 1 exams in January 2015 and we updated it in November 2015.

Firecracker's application has changed significantly since November 2015. With the replacement of Firecracker Legacy with Firecracker MD in September 2015 came an update to Firecracker's daily review algorithm, an overhaul of Firecracker's user interface, and an emphasis on answering USMLE-style clinical case questions. This paper examines users who used Firecracker's updated platform.

We are ready to share some new and updated results now that we have 2016 USMLE Step 1 data.

Summary

- Firecracker MD users score an average of 15 points higher and 3 points higher on the Step 1 exam than non-Firecracker users and Firecracker Legacy users, respectively.
- There are significant correlations between flashcard coverage, effort, and mastery on Firecracker and Step 1 scores. These correlations persist after adjusting for MCAT scores.
- Firecracker's Step 1 practice exam blocks allow for precise student benchmarking. Firecracker's first two 40-question practice exam blocks explain more variance in Step 1 scores than the Comprehensive Basic Sciences Self Assessment exam does.
- Firecracker's practice exam remediation tasks are correlated with significantly increased scores on a subsequent practice exam.

Analysis

We compared the distribution of Firecracker Step 1 scores with the distribution of non-Firecracker Step 1 scores. "Firecrackers" were defined as users who answered at least 1500 flashcards in the Firecracker program prior to their exam. This correlates with at least 1 month of regular Firecracker use. "Non-Firecrackers" were defined as students who submitted Step 1 scores and did not have a Firecracker account or answered fewer than 1500 flashcards in Firecracker prior to when they took their exam.

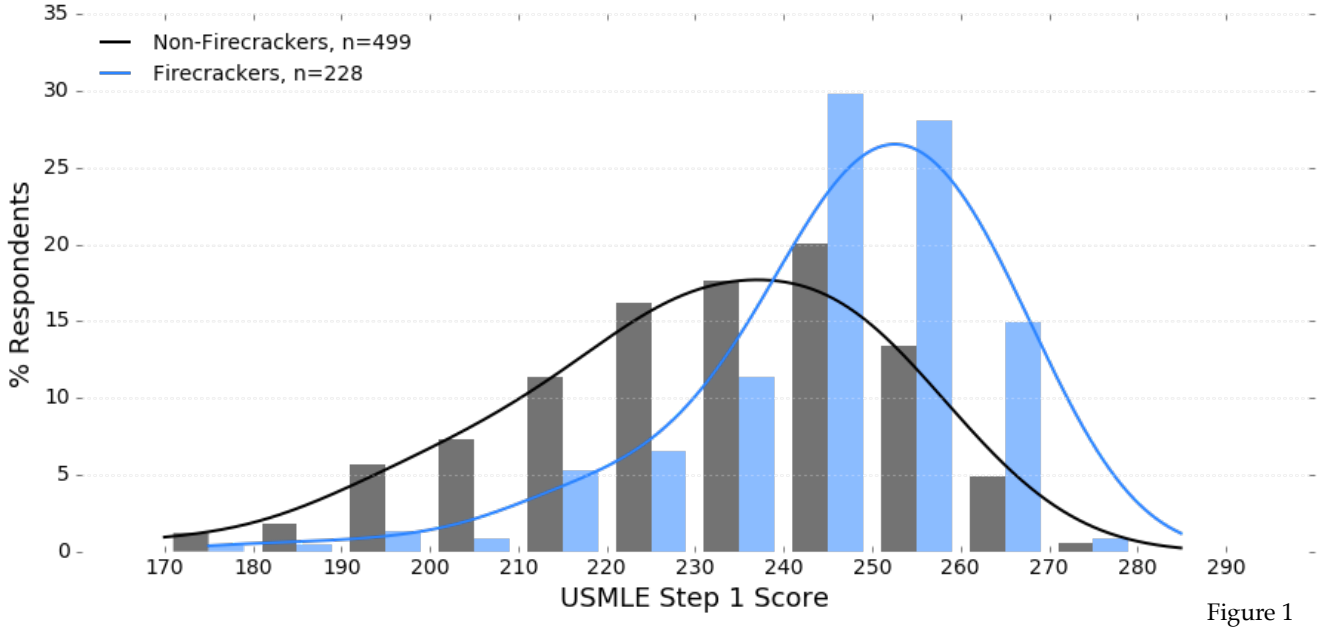


Figure 1

The average non-Firecracker Step 1 score is 229.9 with a standard deviation of 21.3. This is very close to the reported 2015 national statistics[1] (national average of 229 and standard deviation of 20). The average Firecracker score is 245.0 with a standard deviation of 16.6. The distribution of Firecracker scores is 15 points higher on average and there is less variance in Firecracker Step 1 scores than non-Firecracker scores. An unpooled two-sided t-test finds that the average Firecracker score is significantly higher than the average non-Firecracker score ($p < 0.00001$).

[1] http://www.usmle.org/pdfs/transcripts/USMLE_Step_Examination_Score_Interpretation_Guidelines.pdf

We also took a look at how Firecracker MD users performed on their Step 1 exams compared to Firecracker Legacy users. Again, "Firecrackers" were defined as users who answered at least 1500 flashcards in the Firecracker platform.

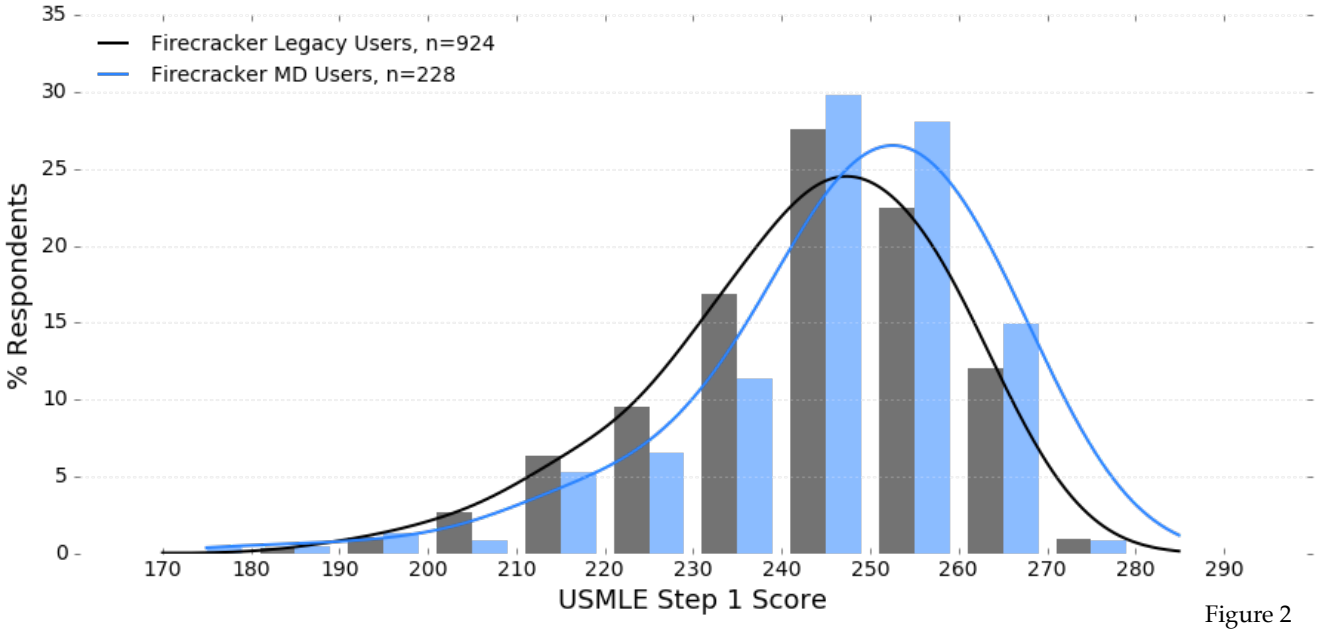
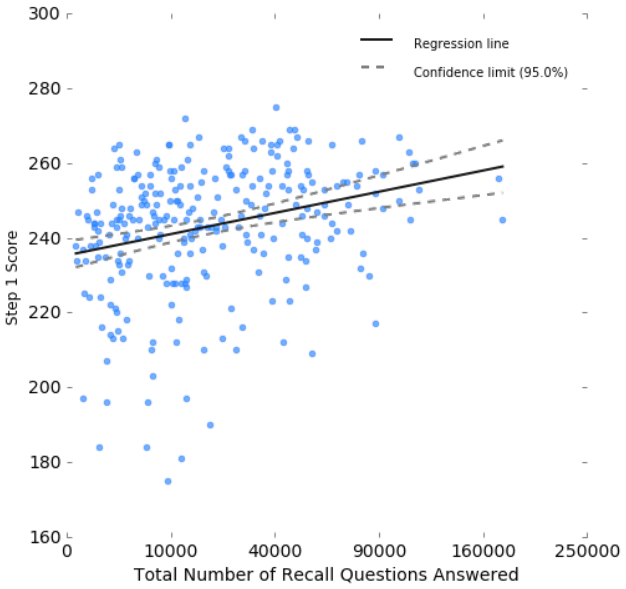


Figure 2

The average Firecracker Legacy Step 1 score is a 242.0 while the average Firecracker MD Step 1 score is a 245.0. The standard deviation in scores is 16.3 and 16.6 for Firecracker Legacy users and Firecracker MD users, respectively. A pooled two sided t-test finds that Firecracker MD users scored significantly higher on their Step 1 exams than Firecracker Legacy users ($p = 0.014$). The discrepancy in sample sizes is due to the fact that we have collected Firecracker Legacy Step 1 scores for three years and Firecracker MD scores for only one year. It is important to note that the population average of Step 1 scores has not changed in the past three years[2].

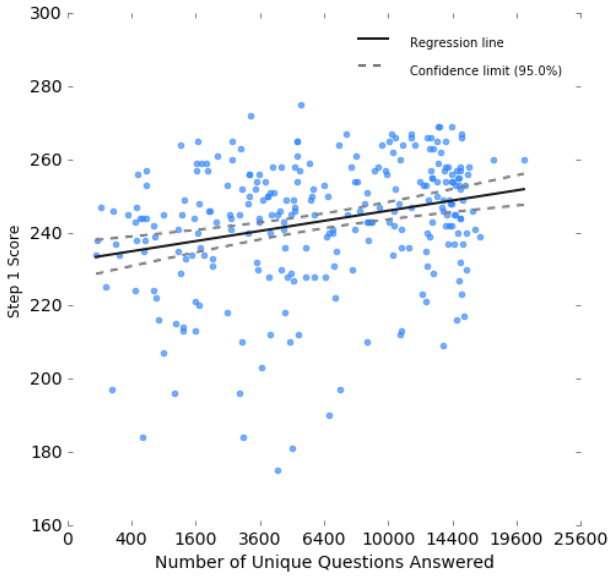
[2] http://www.usmle.org/pdfs/transcripts/USMLE_Step_Examination_Score_Interpretation_Guidelines.pdf

We examined how individual Firecracker usage metrics correlated with performance on the USMLE Step 1 exam.



r=0.285, n=269, p<0.001

r=0.262, n=269, p<0.001



r=0.281, n=269, p<0.001

Clockwise from top left:
Figure 3, Figure 4, Figure 5

We observe significant correlations between the Firecracker usage metrics and USMLE Step 1 scores. Exposure (Figure 3 and 5) to and mastery (Figure 4) of flashcards within the Firecracker platform is correlated with significantly increased Step 1 scores. None of the bivariate correlations shown above are significantly different from the bivariate correlations examined with Firecracker Legacy in previous whitepapers.

MCAT scores are significant predictors of Step 1 scores[3] and can be used as a baseline measurement to adjust for potential selection bias in our sample. We received MCAT scores from 58 of the 228 Firecracker users who submitted Step 1 scores and found that the correlations between Firecracker usage and Step 1 scores persist even after adjusting for MCAT scores. Prior survey data has also shown no significant difference in the number of hours preparing for Step 1 between Firecrackers and non-Firecrackers (Appendix Figure 11).

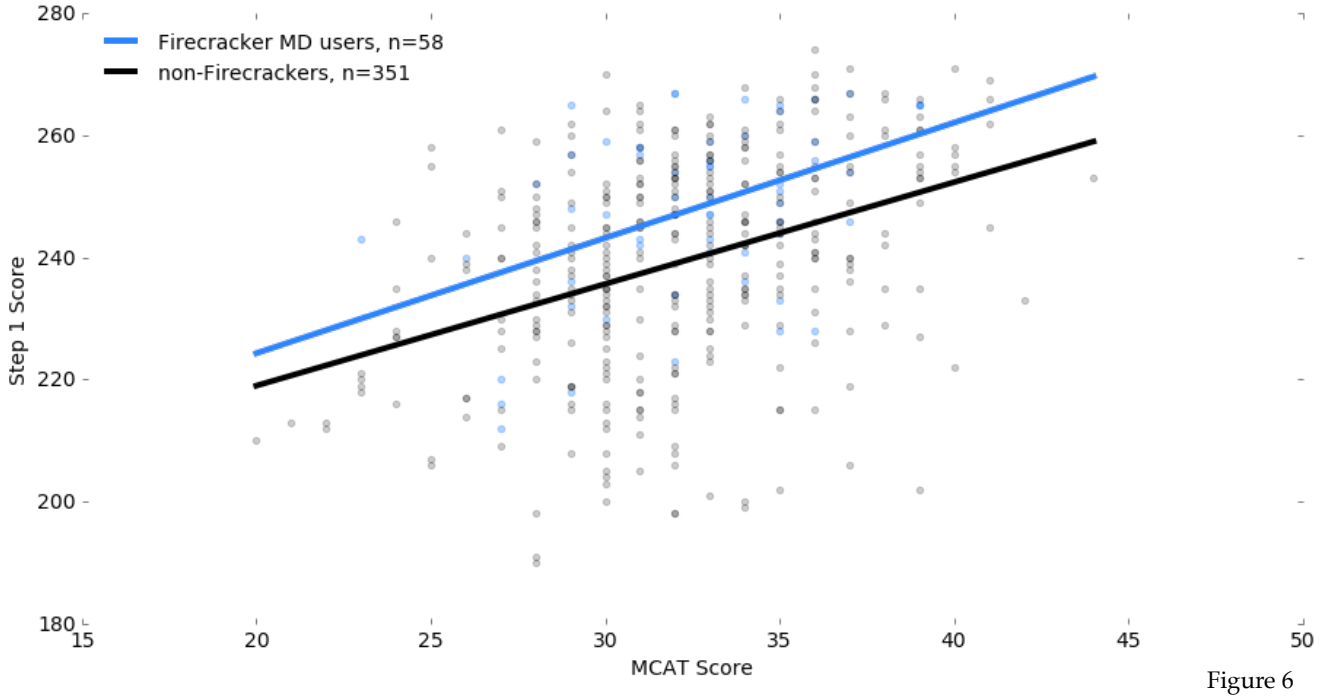


Figure 6

The regression lines in Figure 6 show that Firecracker users generally outperform non-Firecrackers with the same MCAT score by 8 points on the Step 1 exam.

[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.

A multiple regression model gives more detail on the MCAT-adjusted correlation between Firecracker usage and Step 1 scores.

Table 1

OLS Regression Results

Dep. Variable:	**Step 1 Score**	R-squared:	0.469
Model:	OLS	Adj. R-squared:	0.443
No. Observations:	66	AIC:	535.7
Df Model:	3	BIC:	544.5

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	135.2168	15.689	8.619	0.000	103.855	166.579
MCAT Score	2.5144	0.411	6.123	0.000	1.694	3.335
(# of Flashcards Seen)^0.5	0.1079	0.052	2.061	0.044	0.003	0.213
Average Flashcard Score	5.8433	3.071	1.903	0.062	-0.296	11.982

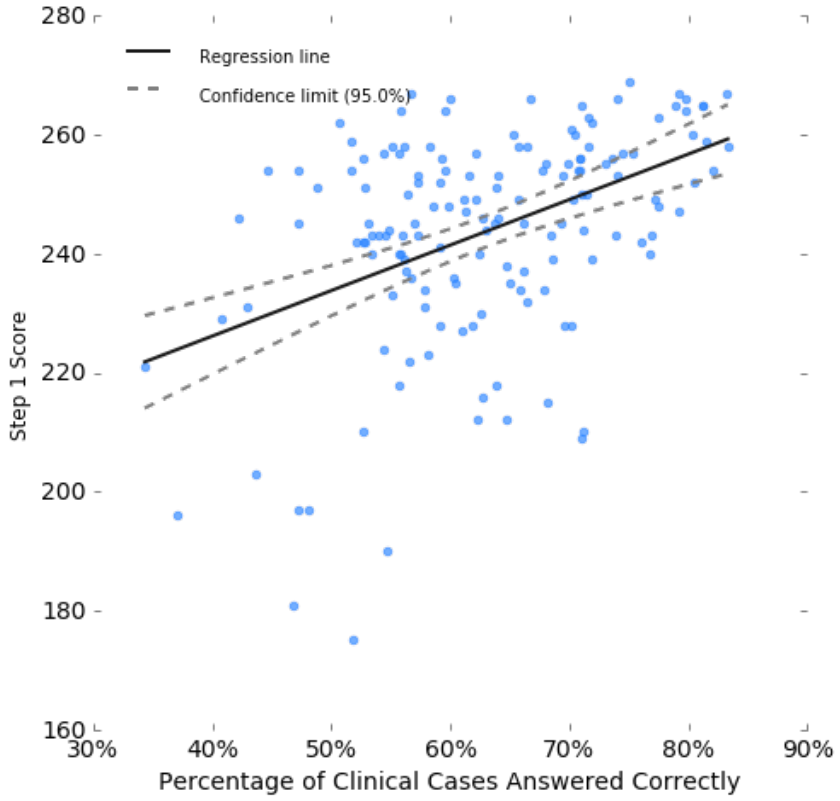
Adding Firecracker usage statistics to an MCAT only model explains an additional 7% of the variance in Step 1 scores (R^2=*0.402* for the MCAT only model, Appendix Table 4). Each regression coefficient is positive and significant meaning that a student's coverage and mastery of Firecracker material are both associated with increased Step 1 scores, even after adjusting for MCAT scores. For example, the model in Table 1 predicts that a student who has answered 1000 flashcards on Firecracker will score nearly 2.5 points higher on their Step 1 exam than someone who has only answered 100 flashcards with the same MCAT score.

Benchmarking and Remediation

With the launch of Firecracker MD, Firecracker revamped its bank of clinical case questions and practice tests. The updated Firecracker MD algorithm puts an emphasis on regularly answering clinical case questions. We took a look at this new data to understand Firecracker's ability to benchmark students and give them effective remediation tasks.

Interestingly, we found no significant correlation between the number of clinical case questions a student attempted in Firecracker and their performance on the Step 1 Exam (Appendix Figure 9). However, there is a very strong correlation between a user's percentage of correct responses on their first attempts of clinical case questions and their Step 1 score[4].

Figure 7



r=0.446, n=151, p<0.001

[4] We filtered for questions answered within a year of when the user took their Step 1 exam and for users who answered at least 40 clinical case questions.

The existence of a correlation between study effort using flashcards and Step 1 scores (Figures 3 and 5) and an the absence of a correlation between number of attempted clinical case questions and Step 1 scores (Appendix Figure 9) are noteworthy. It is possible that simply attempting more clinical case questions does not increase knowledge, while reviewing flashcards on key medical concepts does.

Firecracker's platform uses flashcards in its spaced review system to ensure that students have learned the key medical concepts on the USMLE. Firecracker then uses clinical case questions to test the application of these key medical concepts (Appendix Figure 10). In a multiple regression model, we found that a user's study metrics on flashcards strongly correlated with their subsequent performance on related clinical case questions[5].

Table 2

Logit Regression Results

Dep. Variable:	**Answer Status**	No. Observations:		3759		
Model:	Logit	Df Model:		5		
	coef	std err	z	P>|z|	[95.0% Conf.	Int.]
Intercept	-0.0406	0.038	-1.073	0.283	-0.115	0.034
User Skill	0.5914	0.042	14.226	0.000	0.510	0.673
Question Difficulty	1.0523	0.045	23.214	0.000	0.963	1.141
# of Flashcards Seen	0.1361	0.041	3.359	0.001	0.057	0.216
Average Flashcard Score	0.0800	0.038	2.089	0.037	0.005	0.155
Days Since Last Review	-0.1567	0.039	-4.066	0.000	-0.232	-0.081

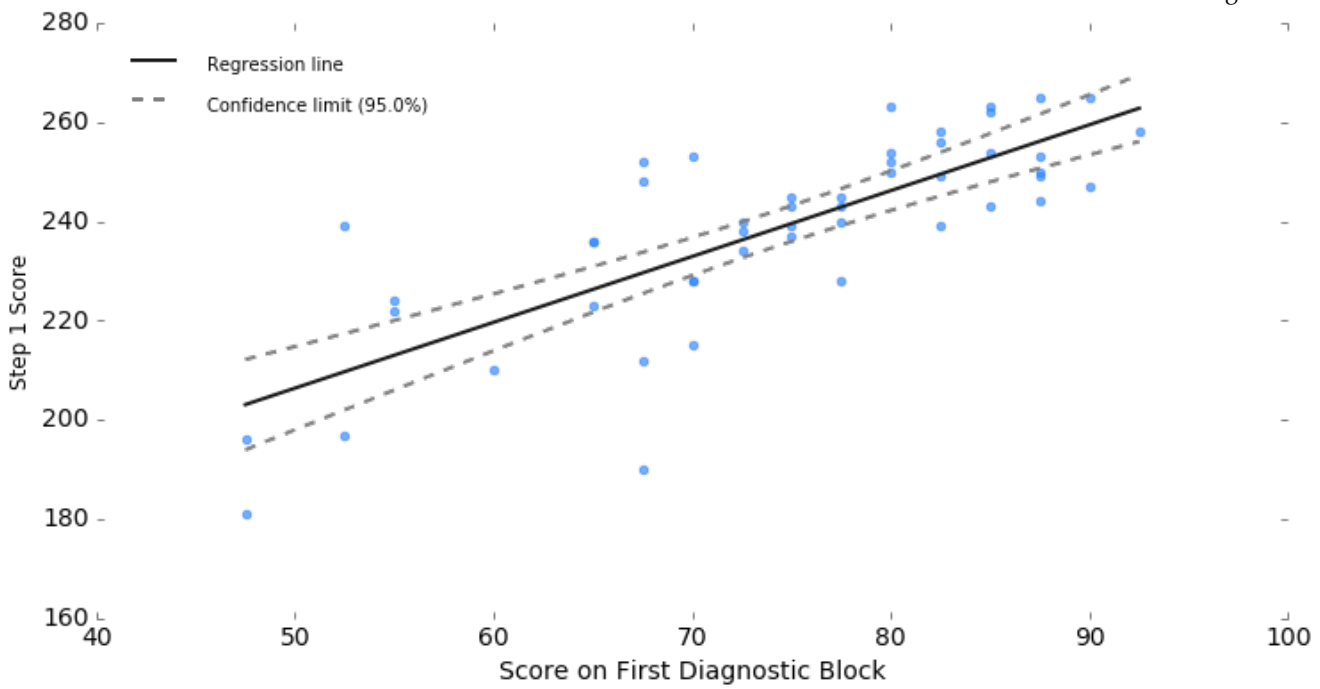
The logistic regression model shown in Table 2 is able to correctly predict a user's response on a clinical case question over 71.4% of the time. The model shows us that, even after adjusting for user skill and question difficulty, consistently answering and mastering flashcards significantly increases a user's chance to answer a related clinical case question correctly. We also see evidence of the "forgetting effect": as more time elapses since the user last studied related flashcards their chance of answering the clinical case question correctly decreases.

Firecracker's algorithm helps users prepare for clinical case questions by pushing them to re-review flashcards with low scores and by resurfacing flashcards over time so that the key concepts stay fresh.

[5] Note that the variables in Table 2 have been standardized using item response theory. "User Skill" is calculated as the proportion of all other clinical cases the user answers correctly, "Question Difficulty" is calculated as the proportion of all other users who answer the clinical case question correctly, and "Days Since Last Review" is calculated as the number of days between when the user last reviewed a related flashcard and when the user answered the clinical case question.

In May 2016, Firecracker launched "Dedicated Test Prep Mode" for users who were within 3 months of taking their board exams. In Dedicated Test Prep Mode, or DTP mode, users receive a weekly 40-question practice exam block. We found that the first exam block in DTP mode alone had a very strong correlation with Step 1 scores (Figure 8). This 40-question practice exam block explained 61% of the variance in Step 1 scores. For reference, the CBSSA, a 200-question exam administered over four hours, explains 67% of the variance in Step 1 scores[6].

Figure 8



r=0.781, n=51, p<0.0001

Adding a second diagnostic block explains 76% of the variance in Step 1 scores (Appendix Table 5) - a 15% increase from just the first diagnostic block and 8% higher than the CBSSA.

[6] http://journals.lww.com/academicmedicine/Abstract/2010/10001/Relationship_Between_Performance_on_the_NBME.26.aspx

Firecracker's system offers targeted remediation tasks for users after they take a practice exam in DTP mode. Remediation tasks depend on how many questions a student answered incorrectly on their first exam and generally range between 200 and 500 flashcards. We found that completion of these remediation tasks is significantly correlated with increased performance on the next practice exam taken in Dedicated Test Prep mode.

Users who did not complete their remediation tasks saw no significant change between their first and second practice exam scores, while users who completed the remediation task saw a significant increase on their second exam score.

Table 3

OLS Regression Results

Dep. Variable:	**Block 2 Score**	R-squared:	0.671
Model:	OLS	Adj. R-squared:	0.668
No. Observations:	251	AIC:	-514.6
Df Model:	2	BIC:	-504.0

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	8.9370	2.539	3.520	0.001	3.937	13.937
Remediation Task Completed	2.9068	1.634	1.779	0.077	-0.312	6.126
Block 1 Score	0.8477	0.038	22.469	0.000	0.773	0.922

The regression results in Table 3 show that we expect a user who completes their remediation task to score almost 3% higher on the followup exam than a user with the same first exam score who didn't complete the remediation task.

Conclusion

We continue to see significant correlations between the extent to which a student uses, covers, and masters Firecracker's material and their performance on the Step 1 exam. Firecracker users don't report studying more than non-Firecrackers, and these correlations persist even after adjusting for MCAT scores as a baseline measurement.

Firecracker helps students apply their knowledge in clinical scenarios. There is a strong correlation between mastery of key medical concepts using Firecracker's flashcards and performance on related clinical case questions.

Firecracker's DTP mode practice exam blocks are highly predictive of Step 1 exam scores and Firecracker's targeted remediation tasks lead to significantly improved scores between exams.

Glossary

Term	Description
Firecracker Legacy	Firecracker's web platform prior to its replacement with Firecracker MD
Recall Question	A flashcard question that requires the user to actively recall a key medical concept
Recall Score	A user's self reported confidence on a recall question (1-5 with 5 being full recall)
Average Recall Score	A student's average recall score on their most recent flashcard reviews
Clinical Case Question	A USMLE-style clinical vignette with an objectively correct answer that tests the user on several key medical concepts
DTP Mode	Firecracker's dedicated test prep mode. In DTP mode students receive weekly 40-question practice exams with targeted remediation tasks

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 ($p<0.001$).

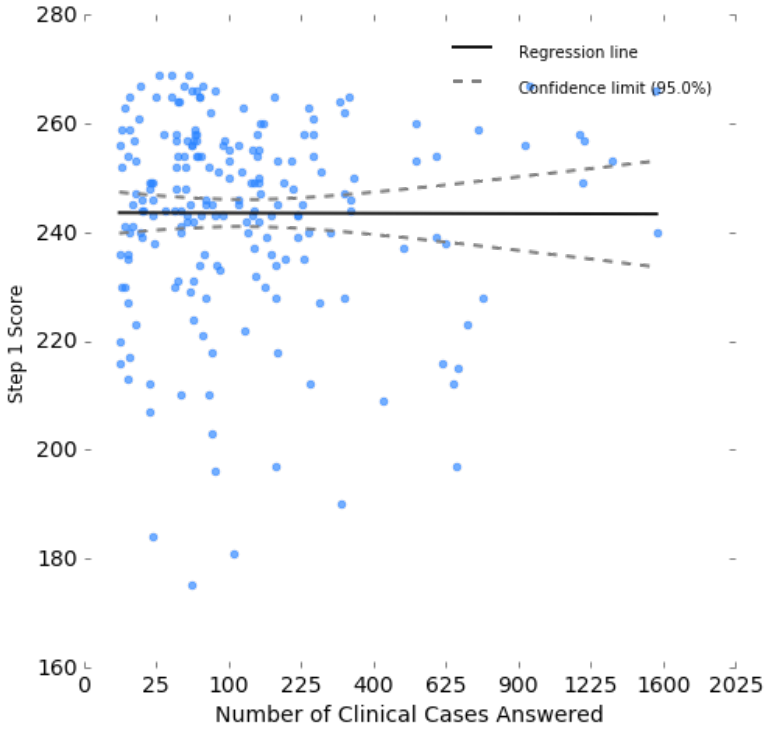
Table 4

OLS Regression Results

Dep. Variable:	**Step 1 Score**	R-squared:			0.402	
No. Observations:	66	Adj. R-squared:			0.393	
	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	156.6353	13.522	11.584	0.000	129.622	183.648
MCAT	2.7516	0.420	6.558	0.000	1.913	3.590

Scatterplot of the number of clinical case questions answered and Step 1 scores. There is no significant correlation.

Figure 9



r=-0.002, n=202, p=0.97

Multiple regression results for first two diagnostic blocks in Firecracker's Dedicated Test Prep mode.

Table 5

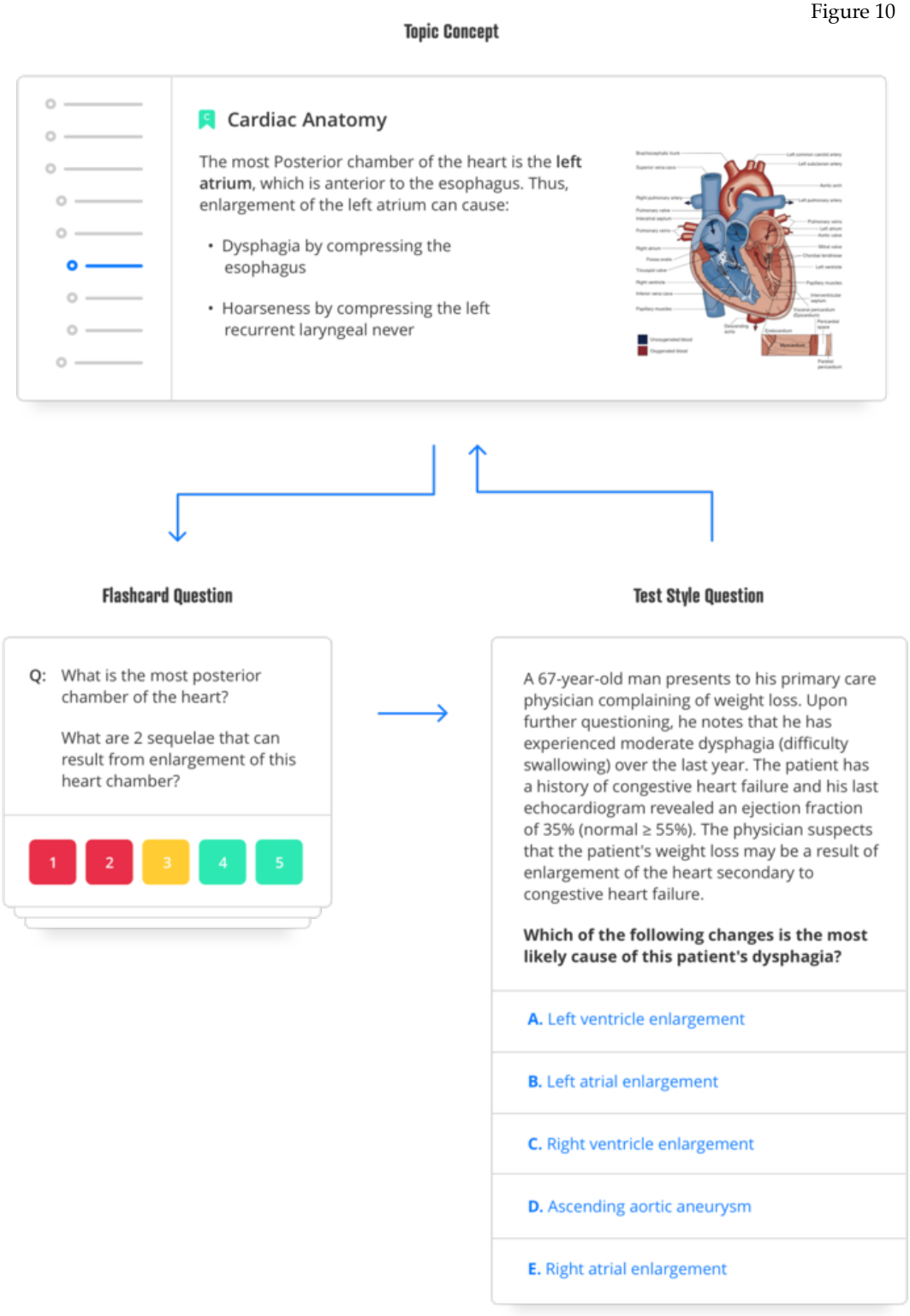
OLS Regression Results

Dep. Variable:	**Step 1 Score**	R-squared:	0.760
Model:	OLS	Adj. R-squared:	0.733
No. Observations:	21	AIC:	154.0
Df Model:	2	BIC:	157.2

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	138.7373	13.315	10.419	0.000	110.763	166.712
Block 2 Score	0.6580	0.212	3.107	0.006	0.213	1.103
Block 1 Score	0.6868	0.216	3.179	0.005	0.233	1.141

Diagram of Firecracker's content model. Users are guided to first review key medical concepts in Firecracker's topics, then master the key medical concepts using flashcard review questions, and ultimately apply the key medical concept in clinical applications.

Figure 10



In our 2015 survey, 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 ($p=0.967$).

Figure 11