Spark Evaluation

Fil Babalievsky and Atishay Sehgal

8/1/2018

Abstract

We evaluate the effect of Project SPARK, an effort to introduce aerobic exercise to Staten Island's K-12 PE classes, on academic outcomes. We find mild evidence that it may have had a positive impact, and recommend further steps.

Introduction:

The Staten Island Borough President's office has conducted a preliminary investigation into the effects of aerobic exercise on academic achievement, in consultation with Dr. John Ratey. This program was inspired by and named for his book "SPARK", an investigation of the benefits of exercise. This was a minimal pilot program intended as a proof of concept, yet it may have had a small positive effect.

Program details:

One of Staten Island's public schools agreed to serve as a small scale pilot for SPARK. They took a group of just under one hundred ninth graders and, starting in the second quarter of the school year, divided them into two groups. The students in two physical education classes, comprising two thirds of the group, replaced their standard sports-based curriculum with intense cardio exercise. At the start of the program, this meant seven minutes of intense work on treadmills, eventually expanded to twenty. Observers reported that, even in the early phase of the program, students were visibly exhausted. Note that both groups of students had the same physical education curriculum in the first quarter, providing a useful baseline.

Also, the students were still subject to the state physical education curriculum.

A typical physical education curriculum focuses more heavily on sports than on getting students' heart rate up. SPARK replaced the games that students typically play with intense workouts designed to keep them at a high percentage of their maximum heart rates.

The students were split based on convenience. Two gym classes were arbitrarily chosen and the third was left out as a control, but the school officials who separated the students did not do so based on any systematic difference between the students.

Research Design and Data:

We were given data on student performance in math, English, and science classes, with one average score per quarter. Our main specification consists of a differences in differences approach. The first quarter, when all students had the same gym curriculum, provides a baseline for comparison. We test if student academic growth differed between treatment and control over the remaining marking periods.

Our first research design compares the difference in grades between the first and subsequent marking periods for SPARK and non SPARK students, without controls for type of class. We run three regressions in this format, with the three dependent variables being the change in grades between the first marking period and the three subsequent marking periods.

The regression design is as follows:

$$\Delta_{i,s,m} = \alpha_m + \beta_m \cdot SPARK_i$$

Here $\Delta_{i,s,m}$ was the change in score for student i in subject s from marking period 1 to marking period m, and $SPARK_i$ is a dummy taking value 1 if and only if the student was in SPARK.

Our second research design considers the possibility that SPARK might have had different effects on different subjects. We therefore add controls for the type of subject and replace the SPARK dummy variable with an interaction term between SPARK and each of the three types of classes.

The design is as follows:

$$\Delta_{i,s,m} = \sum_j \alpha_{j,m} + \sum_j \beta_{j,m} \cdot SPARK_{i,j}$$

Here α_j takes value 1 if $j = s$ and zero otherwise, and $SPARK_{i,j}$ takes value 1 only if the student is in SPARK and if $j = s$.

The levels of j and s correspond to math, English, and science.

As a handful of students were in a higher-level math course, we run one further test where the levels of j and s correspond to English, science, and each of the two math courses.

All outcome variables are based on student grades, which are on the usual 0 to 100 scale. A 2.5 point coefficient, therefore, means that SPARK students saw their scores increase by a quarter of a grade level relative to non-SPARK students.

In both regression designs, the observations are student-classes.

Results:

First we report the simplest output, not broken up by subject. Note that some students did drop out of the sample in the middle of the school year. We do not know why this is the case.

Table 1:

	Dependent variable:		
	MP1 to MP2	MP1 to MP3	MP1 to MP4
	(1)	(2)	(3)
Spark	2.391**	0.547	1.442
	(0.932)	(1.248)	(0.983)
Constant	−6.385***	−8.835***	−6.029***
	(0.733)	(0.973)	(0.767)
Observations	273	263	263
R^2	0.024	0.001	0.008
Adjusted R^2	0.020	−0.003	0.004
Residual Std. Error	7.478 (df = 271)	9.878 (df = 261)	7.781 (df = 261)
F Statistic	6.579** (df = 1; 271)	0.192 (df = 1; 261)	2.151 (df = 1; 261)
Note:			*p<0.1; **p<0.05; ***p<0.01

Next, we break out results by subject.

Table 2:

	Dependent variable:		
	MP1 to MP2	MP1 to MP3	MP1 to MP4
	(1)	(2)	(3)
Spark_math	3.232**	0.278	2.857*
	(1.612)	(2.025)	(1.703)
Spark_sci	2.907*	0.239	0.089
	(1.603)	(1.996)	(1.679)
Spark_eng	0.986	1.396	1.456
	(1.603)	(1.996)	(1.679)
math	2.380	−8.962***	−3.965**
	(1.792)	(2.224)	(1.870)
eng	2.743	−7.629***	−3.914**
	(1.779)	(2.191)	(1.843)
Constant	−8.086***	−3.371**	−3.429***
	(1.258)	(1.549)	(1.303)
Observations	273	263	263
R^2	0.048	0.153	0.041
Adjusted R^2	0.030	0.136	0.023
Residual Std. Error	7.441 (df = 267)	9.165 (df = 257)	7.708 (df = 257)
F Statistic	2.675** (df = 5; 267)	9.282*** (df = 5; 257)	2.225* (df = 5; 257)
Note:			*p<0.1; **p<0.05; ***p<0.01

Finally we break math into two separate categories.

Table 3:

	Dependent variable:		
	MP1 to MP2	MP1 to MP3	MP1 to MP4
	(1)	(2)	(3)
Spark_math	3.449^{**}	-0.675	2.735
	(1.676)	(2.106)	(1.781)
Spark_sci	2.907^{*}	0.239	0.089
	(1.604)	(1.992)	(1.685)
Spark_eng	0.986	1.396	1.456
	(1.604)	(1.992)	(1.685)
Spark_adv_math	-9.449	7.800	2.140
	(8.071)	(9.927)	(8.398)
math	2.086	-8.910^{***}	-3.915^{**}
	(1.806)	(2.237)	(1.892)
eng	2.743	-7.629^{***}	-3.914^{**}
	(1.779)	(2.186)	(1.850)
adv_math	10.000	-1.719	-1.656
	(7.556)	(9.289)	(7.858)
Constant	-8.086^{***}	-3.371^{**}	-3.429^{***}
	(1.258)	(1.546)	(1.308)
Observations	273	263	263
R^2	0.054	0.163	0.042
Adjusted R^2	0.029	0.140	0.015
Residual Std. Error	7.444 (df = 265)	9.147 (df = 255)	7.738 (df = 255)
F Statistic	2.165^{**} (df = 7; 265)	7.092^{***} (df = 7; 255)	1.587 (df = 7; 255)
Note:			*p<0.1; **p<0.05; ***p<0.01

Analysis

The point estimates for math were generally higher than the point estimates for the other subjects, and were more often significant. The greater responsiveness of math scores to an intervention is in line with much of the rest of the literature (see Graff Zivin et al 2018 or the discussion in Fryer 2017).

Most of the point estimates were positive, and a handful were significant. The largest effect size was a roughly one-third grade level difference in math from quarter one to quarter two. That is, students in the treatment group saw a change in grades that was one third of a grade level higher from marking period to marking period than the control. Still, these results were not all that large or significant.

We include an R notebook in the replication page for this project with more robustness tests, none of which meaningfully alter our conclusions.

Recommendations

This minimal pilot program offers mild evidence that replacing students' sports-based curriculum with aerobic activity may improve academic performance. Given that this is just in addition to the more obvious fitness benefits of exercise, we recommend exploring this effect further. The ideal research design would be a pre-registered randomized trial, publicly outlining the empirical strategy prior to implementing it.

We suggest that every gym class in Staten Island be placed into a treatment and control group at random, with the treatment group switching from a sports-based curriculum to an aerobic curriculum.

We also recommend collecting baseline data on each of the students, especially academic performance in the prior year and level of fitness going into the school year, and committing in the pre-registration plan to divide students along these pre-selected variables to test for heterogeneous effects. This will help us learn whether less fit or less academically successful students benefit more from this intervention.

Lastly, we recommend that the replication file with all the codes needed to produce these analyses be made public. This will enhance our credibility and allow others to understand and perhaps suggest improvements for our analysis. The data, which includes identifying information, will not be made available. In our output files, we also anonymize all identifying information.

Citations

Beautrelet, I. (2016, December 13). Clustered Standard Errors in R [Web log post]. Retrieved August 2, 2018, from https://economictheoryblog.com/2016/12/13/clustered-standard-errors-in-r/

Fryer Jr, R. G. (2017). The production of human capital in developed countries: Evidence from 196 randomized field experimentsa. In Handbook of Economic Field Experiments (Vol. 2, pp. 95-322). North-Holland.

Graff Zivin, J., Hsiang, S. M., & Neidell, M. (2018). Temperature and human capital in the short and long run. Journal of the Association of Environmental and Resource Economists, 5(1), 77-105.

Hendricks, Paul (2015). anonymizer: Anonymize Data Containing Personally Identifiable Information. R package version 0.2.0. https://cran.r-project.org/web/packages/anonymizer/index.html

Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.2. https://CRAN.R-project.org/package=stargazer