Statistical report for Sportsometry

Walter Dempsey[*]

December 30, 2019

Abstract

1 Overview

We detail the statistical results for the analysis of test outcomes administered to students who participated in the Sportsometry program as well as a set of students who did not participate in the program. Data analyzed included student's school, gender, grade, pre-test score, pre-test total (i.e., max score of pre-test), post-test score, post-test total. In this write-up, we focus exclusively on these summative assessments administered at the beginning and end of the program. We are interested in assessing whether the difference in performance is statistically significantly larger for those students in Sporstometry than those in the control group. Using the data received, we reach the following general conclusions:

1. We see a statistically significant improvement from pre- to post-test for students in the Sportsometry population on average across schools.

2. Comparing only those students at BTW, we see the improvement for students in the Sportsometry program is statistically significantly larger than the improvement for students who did not take the program. This improvement was robust to sensitivity analysis.

2 Data description

There are 383 total observations of students who took the summative assessments. This consists of 226 BTW, 65 no school assignment, 39 WRSA, 15 Clemente, 3 Wintergreen, 3 Clinton Ave, 3 Nathan Hale, 2 Church Street, 2 Elm City, 2 Ross Woodward, and 1 student at Amistad, Barnard, Bishop Woods, Columbus, Conte West Hills, ESUMS, Ferra East Haven, Highville, LBCS, Morrow-Sheriden, Quinnipiac, Ridge Road, Savin Rock, St. Lawrence, St. Rita, Strong, Troup, West Woods, and Wexler Grant. The number of questions on each test ranges from 5 to 11. There was some missing data and for some of the analysis done, the observations with missing requisite data were disregarded. For covariates such as gender and grade, we assume this missingness is random (i.e., does not depend on

[*]Department of Statistics, Harvard University, E-mail: wdempsey@fas.harvard.edu

the true values). For outcomes such as pre- or post-test score, we assume that the missingness is conditionally independent of the potential score given the covariates. If missingness did not satisfy these conditions, we would have biased results and would require additional sensitivity analysis to be performed.

3 Data analysis

3.1 Restriction to BTW students

For each BTW individual, we compute the fraction of correctly answered questions for the pre- and post-test. We then take the difference between these two fractions. The boxplot below displays this improvement split between those individuals who were in the Sportsometry program and those who were not in the program (i.e., "Control"). We can see a difference of 0.125 between the median for the control and Sportsometry groups. Moreover, comparing means we see that the Sportsometry students on average outperform the students in the control group by 0.116.

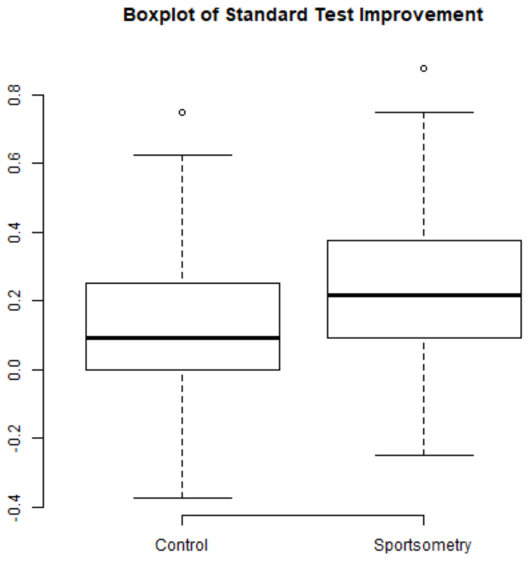

Figure 1: Boxplot of standard test improvement for control and Sportsometry groups

When we compare two samples to see if the difference in means is statistically significant, we must ensure that the underlying assumptions hold; namely, that the two populations are identical to one another in underlying characteristics. In this case, we want to check that the two samples are similar in Gender and Grade. The Gender distribution is 45 Female to 57 Male for the control group and 33 Female to 39 Male in the Sportsometry group. The Grade distribution is, $(0, 0, 32, 34, 0, 18, 0, 0, 0, 11, 7)$ for grades K through 10 for the control group and $(0, 0, 23, 24, 6, 9, 1, 0, 0, 4, 5)$ for grades K through 10 for the Sportsometry group. These differences are small but still may make the standard two-sample test results suspect.

Therefore, we will peform sensitivity analysis using inverse-probability weighting to adjust for potential differences.

We compute a two-sample pooled variance test with the following means and pooled sample standard deviations:

$$\begin{aligned}\bar{x}_{\text{SP}} &= 0.226,\\ \bar{x}_{\text{CTRL}} &= 0.110, s = 0.209\\ t^\star &= \frac{0.226 - (0.110)}{0.209 \cdot \sqrt{1/72 + 1/102}}\\ &= 3.60 < t_{0.95,172} = 1.65.\end{aligned}$$

Therefore, we can reject the null hypothesis that there is no difference in improvement across the treatment and control groups. This suggests that there is a statistically significant larger improvement for children who were in the Sportsometry. We also perform a weighted regression where the weights depend on the probability of being in the Sportsometry group given Gender and Grade. This adjusts for the potential imbalance across the treatment and control groups. We have a treatment effect of 0.121 with standard error 0.03 and p-value 9.30×10^{-5}. This suggests the statistically significance is robust to sensitivity analyses of this type.

We also performed a regression analysis to give a more detailed analyis of the data. To do this, we used the following statistical model:

$$\text{Impr.Score}_{ijkl} = \beta_0 + \text{Grade}_j + \text{Treatment}_k + \text{Gender}_l + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$ is a random white-noise term. We include the weights to adjust for the issues regarding potential selection bias.

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	0.1199	0.0601	1.99	0.0477
Grade == 4	0.0693	0.0606	1.14	0.2548
Grade == 5	-0.0124	0.0602	-0.21	0.8366
Grade == 6	-0.2153	0.1333	-1.61	0.1083
Grade == 7	-0.0811	0.0656	-1.24	0.2182
Grade == 8	0.1622	0.2991	0.54	0.5883
Grade == 11	0.0478	0.0733	0.65	0.5150
Txt == ‘Sportsometry’	0.1279	0.0297	4.31	0.0000
Gender == ’Female’	-0.0350	0.0295	-1.19	0.2373

The benefit of regression is that it tries to control for school and gender while testing if Sportsometry students see a positive impact to their percent improvement in scores. Using above, we say that while holding everything else constant, moving a student from the control group to the Sportsometry group results in an 0.128 increase in the percent difference between pre- and post-test scores. Here, we have a p-value of 2.76×10^{-5}, which suggests the difference is statistically significant. The p-value for the unweighted regression is 4.22×10^{-5}. Thus the weighted regression is more conservative and still finds statistical significance.

3.1.1 Logistic regression

Here, we assess the same data using logistic regression. This regression method accounts for the fact that these are test items with binary outcomes (i.e., right or wrong). We model the test as a binomial distribution. That is, the test is N items (say 6 as in the WRSA case) and the probability of getting a question right is p. Then the probability of getting k out of N items correct is given by

$$\mathrm{pr}(Y = k; N) = \binom{N}{k} p^k (1-p)^{N-k}.$$

We now model the probability p as a function of

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \mathrm{Grade}_j + \mathrm{Treatment}_k + \mathrm{Gender}_l + \epsilon_i.$$

Here, since we are no longer modeling differences, we introduce a *null* level of treatment for both control and Sportsometry kids. This accounts for the pre-test being the same for both groups (i.e., conditional on all other variables there is no difference at baseline between the two groups). To test if the improvement for the Sportsometry group is statistically significantly larger than that for the control group, we fit the same model where the coefficients for "Control" and "Sportsometry" are equal. Twice the difference in the deviance is approximately χ^2 with 1 degree of freedom. Thus the p-value is 4.63×10^{-5}. This supports the conclusions from the prior section. We again perform inverse-probability weighting to adjust for potential differences in treatment and control groups. We find that the conclusions are the same with a p-value of 1.18×10^{-4}.

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-0.7885	0.5394	-1.46	0.1438
Txt == 'Control'	0.1854	0.0640	2.90	0.0038
Txt == 'Sportsometry'	0.4112	0.0733	5.61	0.0000
Gender == 'Female'	-0.2351	0.3970	-0.59	0.5538
Gender == 'Male'	-0.2269	0.3968	-0.57	0.5674
Grade == 2	0.3099	0.6678	0.46	0.6425
Grade == 3	0.4213	0.6717	0.63	0.5305
Grade == 4	0.6909	0.6861	1.01	0.3139
Grade == 5	0.5048	0.6730	0.75	0.4532
Grade == 6	0.5785	0.7682	0.75	0.4514
Grade == 9	0.0761	0.6768	0.11	0.9105
Grade == 10	-0.1024	0.6775	-0.15	0.8799

3.2 Across all data

For all individuals, we compute the fraction of correctly answered questions for the pre- and post-test. We then take the difference between these two fractions. The boxplot below displays this improvement split between those individuals who were in the Sportsometry program and those who were not in the program (i.e., "Control"). We can see a difference of 0.1091 between the median for the control and Sportsometry groups. Moreover, comparing

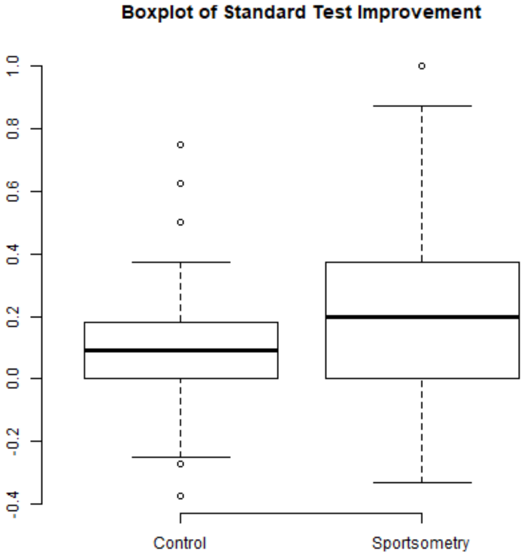

Figure 2: Boxplot of standard test improvement for control and Sportsometry groups

means we see that the Sportsometry students on average outperform the students in the control group by 0.1161.

When we compare two samples to see if the difference in means is statistically significant, we must ensure that the underlying assumptions hold; namely, that the two populations are identical to one another in underlying characteristics. In this case, we want to check that the two samples are similar in Gender and Grade. The Gender distribution is 65 Female to 71 Male for the control group and 60 Female to 77 Male in the Sportsometry group. The Grade distribution is, $(0, 0, 36, 41, 13, 28, 1, 0, 0, 11, 7)$ for grades K through 10 for the control group and $(0, 0, 30, 33, 25, 17, 5, 1, 1, 4, 5)$ for grades K through 10 for the Sportsometry group. Note similar distribution of grades across the two groups.

We compute a two-sample pooled variance test with the following means and pooled sample standard deviations:

$$\begin{aligned}
\bar{x}_{\mathrm{SP}} &= 0.226, \\
\bar{x}_{\mathrm{CTRL}} &= 0.110, s = 0.236 \\
t^{\star} &= \frac{0.226 - 0.110}{0.236 \cdot \sqrt{1/137 + 1/137}} \\
&= 4.077 < t_{0.95,46} = 1.65.
\end{aligned}$$

Therefore, we reject the null hypothesis that there is no difference in improvement across the treatment and control groups. We also perform a weighted regression where the weights depend on the probability of being in the Sportsometry group given Gender and Grade. This adjusts for the potential imbalance across the treatment and control groups. We have a treatment effect of 0.116 with standard error 0.03 and p-value 1.96×10^{-5}. This suggests the lack of statistical significance is robust to sensitivity analyses of this type.

We also performed a regression analysis to give a more detailed analyis of the data. To do this, we used the following statistical model:

$$\text{Impr.Score}_{ijkl} = \beta_0 + \text{Grade}_j + \text{Treatment}_k + \text{Gender}_l + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$ is a random white-noise term. We include the weights to adjust for the issues regarding potential selection bias. The benefit of regression is that it tries to control for school and gender while testing if Sportsometry students see a positive impact to their percent improvement in scores. Using above, we say that while holding everything else constant, moving a student from the control group to the Sportsometry group results in an 0.104 increase in the percent difference between pre- and post-test scores. Here, we have a p-value of 6.27×10^{-5}, which suggests the difference is statistically significant at the 5% level. The p-value for the unweighted regression is 9.79×10^{-5}.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.5408	0.2975	1.82	0.0703
Txt = ‘Sportsometry’	0.1042	0.0256	4.07	0.0001
Gender == ‘Female’	-0.1590	0.2890	-0.55	0.5828
Gender == ‘Male’	-0.1809	0.2890	-0.63	0.5318
Grade == 3	-0.2640	0.0938	-2.81	0.0053
Grade == 4	-0.1658	0.0772	-2.15	0.0327
Grade == 5	-0.2826	0.0767	-3.68	0.0003
Grade == 6	-0.3097	0.0801	-3.87	0.0001
Grade == 7	-0.3751	0.0790	-4.75	0.0000
Grade == 8	-0.3604	0.1246	-2.89	0.0041
Grade == 9	-0.3193	0.2965	-1.08	0.2824
Grade == 10	-0.4640	0.2965	-1.57	0.1188
Grade == 11	-0.2090	0.0899	-2.32	0.0208

3.2.1 Logistic regression

Here, we assess the same data using logistic regression. This regression method accounts for the fact that these are test items with binary outcomes (i.e., right or wrong). We again model the test as a binomial distribution, with the probability p as a function of

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \text{Grade}_j + \text{Treatment}_k + \text{Gender}_l + \epsilon_i.$$

Here, since we are no longer modeling differences, we introduce a *null* level of treatment for both control and Sportsometry kids. This accounts for the pre-test being the same for both groups (i.e., conditional on all other variables there is no difference at baseline between the two groups). To test if the improvement for the Sportsometry group is statistically significantly larger than that for the control group, we fit the same model where the coefficients for “CTRL” and and “SP” are equal. Twice the difference in the deviance is approximately χ^2 with 1 degree of freedom. Thus the p-value is 8.52×10^{-7}. This supports the conclusions from the prior section. We again perform inverse-probability weighting to adjust for potential differences in treatment and control groups. We find that the conclusions are differente with a p-value of 8.52×10^{-9}.

4 Possible statistical improvements

Both the two-sample pooled variance t-test, regression analysis of differences, and the logistic regression analysis support the conclusion that the Sportsometry program leads to a larger improvement in scores than those in the control group.

Concerns about differences in composition of treatment and control groups were dealt with via inverse-probability weighting to create more similar “pseudo-populations”. This technique still showed a statistically significant difference when comparing the difference in fraction of correct answers. Given the size of the dataset, the results are very promising.

To overcome this in the future, it is strongly recommended that there be an extra emphasis on experimental design, if possible. In particular, the method for choosing members of the control group can be made more robust. There are two ways to achieve this: (1) randomization of individuals to Sportsometry/Control. This method may not be possible. Instead, a “propensity score” matching method may be more suitable. Here, for each student in the Sportsometry group, we would aim to find a student with similar characteristics and place use this person as their comparator in the control group. Either way, finding control data on students at various schools will help improve generalizability of results.

Finally, we analyzed the impact of the Sportsometry program across schools. We see that the effect of the program is significant in improving test scores. This may be confounded, of course, with general schooling (and thus the need for a larger control group to account for this issue).

Analytic results: an overview

Here we summarize the key findings of the statistical report for a general audience.

1. **Sportsometry improves test performance.** Students who participated in Sportsometry saw a larger improvement from pre- to post-test score when compared with students who did not participate in Sportsometry.

2. **Students who did not take Sportsometry saw a small improvement in their ability to get each post-test question correct.** This is in line with improvements expected from students learning in school. The improvement in ability was *markedly larger* for students who took Sportsometry.

3. **The effect of Sportsometry on student test scores replicated across school districts.** When running analyses that controlled for school district, the effect of Sportsometry on student performance persisted. This suggests Sporstometry supplements each school's teaching, improving student performance on a relative scale.

4. **The effect of Sportsometry on student test scores replicated across school grade.** When running analyses that controlled for school grade, the effect of Sportsometry on student performance persisted. This suggests Sporstometry supplements teaching per grade level, improving student performance on a relative scale.

5. **The effect of Sportsometry on student test scores replicated across school gender.** When running analyses that controlled for gender, the effect of Sportsometry on student performance persisted. This suggests Sporstometry improves all student performances, not a particular cohort.

6. **The analysis was robust to sensitivity analysis.** A common concern is that analytic results will not replicate when you control for how student's were selected into the Sportsometry program. For example, perhaps younger students were more likely to join Sportsometry. Such students may be expected to experience more rapid increase in performance over a calendar year. Therefore, the effect we estimate would be due to student selection and not the impact of the Sportsometry program. To try and assess this issue, we performed a variety of sensitivity analyses and saw that the results maintain even when controlling for the issue of student selection.

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-1.1072	0.1706	-6.49	0.0000
Txt == ‘Control’	0.1722	0.0566	3.04	0.0023
Txt == ‘Sportsometry’	0.4062	0.0588	6.90	0.0000
Gender == ‘Female’	-0.0869	0.2048	-0.42	0.6713
Gender == ‘Male’	-0.1083	0.2048	-0.53	0.5970
Grade == K	0.0799	0.7303	0.11	0.9129
Grade == 1	-0.1025	0.5628	-0.18	0.8555
Grade == 2	0.2106	0.1733	1.22	0.2242
Grade == 3	0.3227	0.1703	1.90	0.0581
Grade == 4	0.4310	0.1728	2.49	0.0126
Grade == 5	0.4725	0.1681	2.81	0.0049
Grade == 6	0.5141	0.2747	1.87	0.0613
Grade == 7	0.8942	0.4404	2.03	0.0423
Grade == 8	1.0124	0.4321	2.34	0.0191
Grade == 9	-0.0122	0.1935	-0.06	0.9496
Grade == 10	-0.1887	0.1952	-0.97	0.3336
School == ‘Amistad’	0.4659	0.6544	0.71	0.4765
School == ‘Barnard’	0.3572	0.4439	0.80	0.4210
School == ‘Bishop Woods’	0.1514	0.4746	0.32	0.7498
School == ‘BTW’	0.2836	0.1039	2.73	0.0063
School == ‘Church Street’	-0.4795	0.3557	-1.35	0.1777
School == ‘Clemente’	0.0595	0.1669	0.36	0.7214
School == ‘Clinton Ave’	0.2505	0.2618	0.96	0.3387
School == ‘Columbus’	0.6365	0.3803	1.67	0.0942
School === ‘Conte West Hills’	0.5096	0.3812	1.34	0.1813
School == ‘Elm City’	0.4389	0.3318	1.32	0.1860
School == ‘Ferra East Haven’	0.4192	0.4857	0.86	0.3881
School == ‘Highville’	-0.0917	0.5840	-0.16	0.8752
School == ‘Jepson’	-0.1102	0.2926	-0.38	0.7066
School == ‘LBCS’	-2.0891	1.0382	-2.01	0.0442
School == ‘Morrow-Sheriden’	0.4983	0.4315	1.15	0.2482
School == ‘Nathan Hale’	0.6185	0.2352	2.63	0.0086
School == ‘Quinnipiac’	0.3473	0.3952	0.88	0.3796
School == ‘Ridge Road’	-0.2798	0.4712	-0.59	0.5526
School == ‘Ross Woodward’	0.5085	0.2827	1.80	0.0721
School == ‘Savin Rock’	-0.0456	0.4326	-0.11	0.9161
School == ‘St. Lawrence’	0.4659	0.6544	0.71	0.4765
School == ‘Strong’	-1.5363	0.7613	-2.02	0.0436
School == ‘Troup’	-0.2379	0.6865	-0.35	0.7289
School == ‘West Woods’	-0.1588	0.5125	-0.31	0.7567
School == ‘Wexler Grant’	0.1996	0.7169	0.28	0.7807
School == ‘Wintergreen’	-0.3188	0.2898	-1.10	0.2714
School == ‘WRSA’	0.0363	0.1287	0.28	0.7779

Table 1: Logistic regression using all individual data. Assessing the effect of sportsometry while controling for gender, grade, and school.