Machine learning based app for self-evaluation of teacher-specific instructional style and tools

Supplementary materials on further data analysis.

Here we examine non-experimental data collected in the undergraduate course in ordinary differential equations taught by the first author in 2015 and conduct some statistical tests.

We explore various ways of defining the group of students who preferred homework to clickers and the group of students who preferred clickers to homework. We show that, regadless of the exact way of defining these two groups, students who preferred clickers to homework achieved higher exam scores than students who preferred homework to clickers, when controlled for their pre-knowledge measured by exam scores in previous courses.

We also show that, in order to obtain smaller p-values and hence results that are statistically significant, we could have manipulated certain parameters and chosen the "right" statistical tests. For instance, the standard one-sided t-test shows that the null hypothesis that students who prefer clickers to homework and students who prefer homework to clickers achieve same mean exam scores should be rejected with statistical significance $p < 0.1$. However, we also show that our data do not follow the normal distribution and hence the t-test is hardly applicable.

Data

The first author taught an ordinary differential equations course in 2015 and 2016. Students could choose between two learning activities, clickers and handwritten homework (HW). To get the full score for the learning activities, they needed to have solved more than a certain number of exercises of either type. They could invest all their effort into homework and occasionally attend clicker sessions, or they could invest all their effort into clicker sessions and do homework occasionally, or they could try an even mixture of homework and clickers. Clicker sessions were early in the morning and, presumably, low-motivated students preferred homework. Most students had not been previously exposed to clickers but had been familiar to homework since childhood. Students were encouraged but not forced to collaborate on both activities. Either homework or clicker questions covered about 70% of material that later appeared in the final exam, but students did not know this in advance. Below is a sample of the data (names are fake and exam scores have been rescaled so that the minimum is 0 and maximum 100).

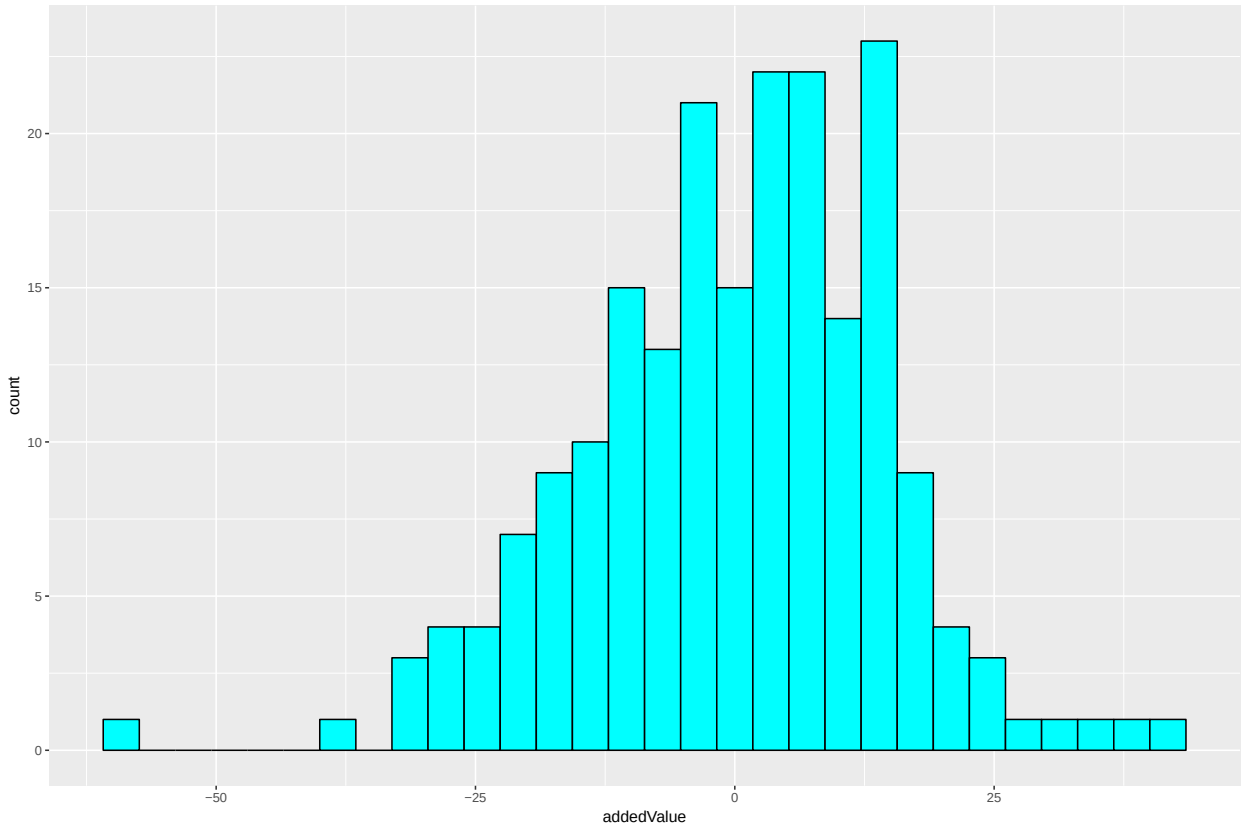
```
##                                name examCalc1 examCalc2 examCalc3 examLA1
## 59              Archmaester Perestan     21.62     45.57     48.89   43.62
## 161        Margot Peake (Lannister)      1.35     21.52     14.44   39.36
## 84                     Jocelyn Swyft     81.08     56.96     60.00   44.68
## 179                       Dorea Sand     56.76     34.18     30.00   60.64
## 190                Lord Philip Plumm     68.92     44.30     20.00   50.00
## 10                   Maggy, The Frog     79.73     73.42     91.11   78.72
## 106              Gormond Goodbrother     89.19     59.49     51.11   31.91
## 177               Archmaester Ebrose     67.57     45.57     47.78   72.34
## 109          Lanna Jast (Lannister)     40.54     35.44     12.22   41.49
## 90  Luthor Tyrell, Son of Theodore     72.97     70.89     34.44   55.32
##     examLA2 writtenHW clickers  exam predictedExamScore addedValue
## 59       45        54       68 76.25           48.55647  27.693534
## 161      20        39       16 46.25           36.61197   9.638033
## 84       65        43       64 71.25           61.02917  10.220828
## 179      55        53       53 38.75           53.29045 -14.540450
```

```
## 190      26        54       41 32.50          46.70915 -14.209147
## 10       82        26       66 95.00          78.10551  16.894492
## 106      37        46       70 65.00          56.36870   8.631302
## 177      40        51       35 40.00          48.20704  -8.207036
## 109      35        55       40 63.75          42.43739  21.312615
## 90       70        46       70 68.75          66.11619   2.633812
```

Here, examCalc1, examCalc2, examCalc3, examLA1, examLA2 are exam scores from previous courses that have been used to predict the final exam score in ordinary differential equations and addedValue is the residual of the model, i.e., actual exam score minus the predicted exam score. Thus addedValue measures students' achievement controlled for the pre-knowledge measured as exam scores in previous courses. The last two columns, predicted exam scores and added values, have been computed in Mathematica via symbolic regression using our app that can be found here: http://tinyurl.com/edex-custom-notebook.

The linear correlation between the actual exam score and the predicted score is 0.7230345. It can be interpreted as the model's precision, share of variance of the response variable explained by predictors, i.e., by previous exams in similar courses.

Histogram of added values



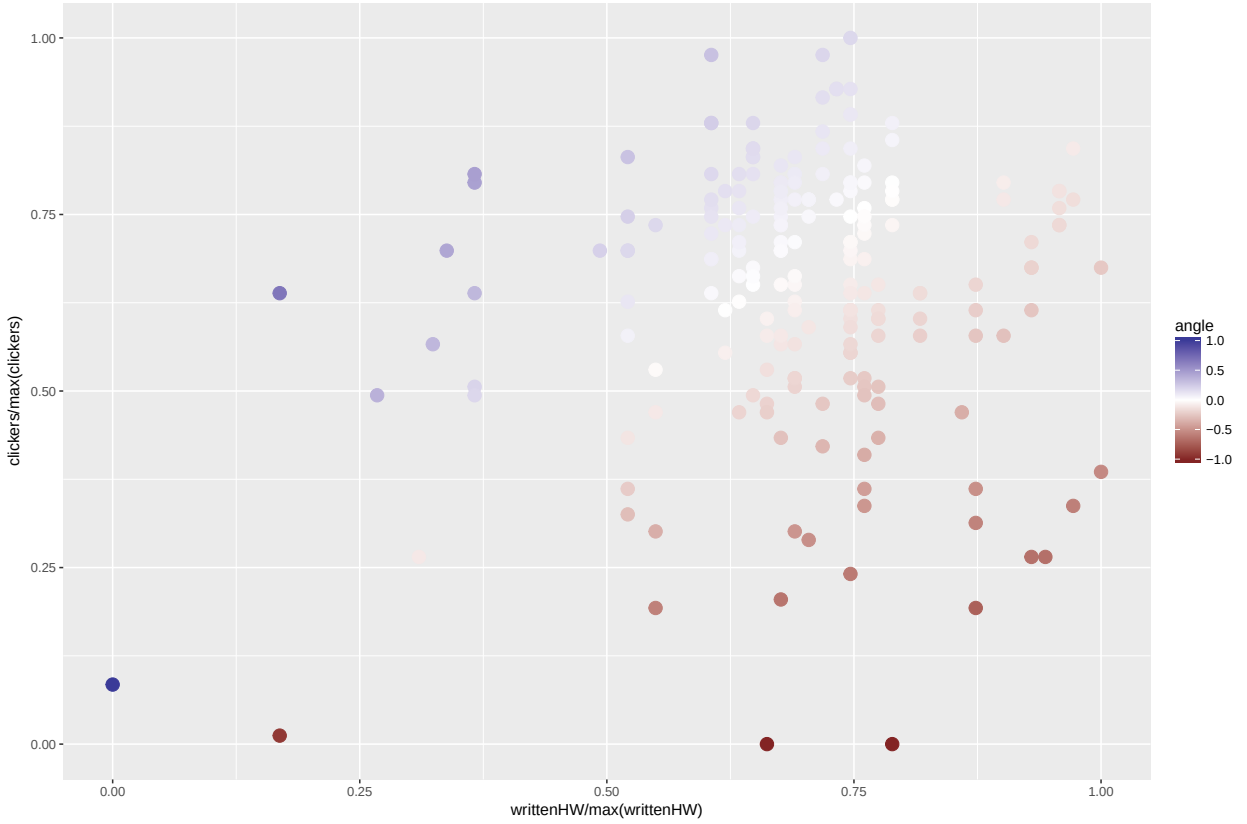
We see that added values do not seem to follow the normal distribution. To confirm this observation, we conduct Shapiro test for normality:

```
##
##  Shapiro-Wilk normality test
##
## data:  X$addedValue
## W = 0.987, p-value = 0.05765
```

A low p-value tells us that it is very unlikely that added values follow a normal distribution.

Scatterplot of clicker scores vs HW scores

Below is a scatterplot where each point represents a student, x-coordinate her HW score and y-coordinate her clicker score. Both HW and clicker scores have been rescaled so that the minimum is 0 and maximum is 1. The color reporesents preference towards HW (red, bottom/right corner of the diagram) or clickers (blue, top/left corner of the diagram). With an appropriate level of committment, it wasn't too hard for students to solve a large number of clicker questions or a large number of HW questions and hence x and y coordinates in this plot measure preference towards one or the other learning activity rather than a student's talent.



Statistical hypothesis testing

Here, we identify

- Group 1 (HW-lovers) - students who prefer homework to clickers
- Group 2 (clicker-lovers) - students who prefer clickers to homework

Then we test the null hypothesis that their added values are, on the average (i.e., mean or median added value) equal vs the alternative hypothesis that

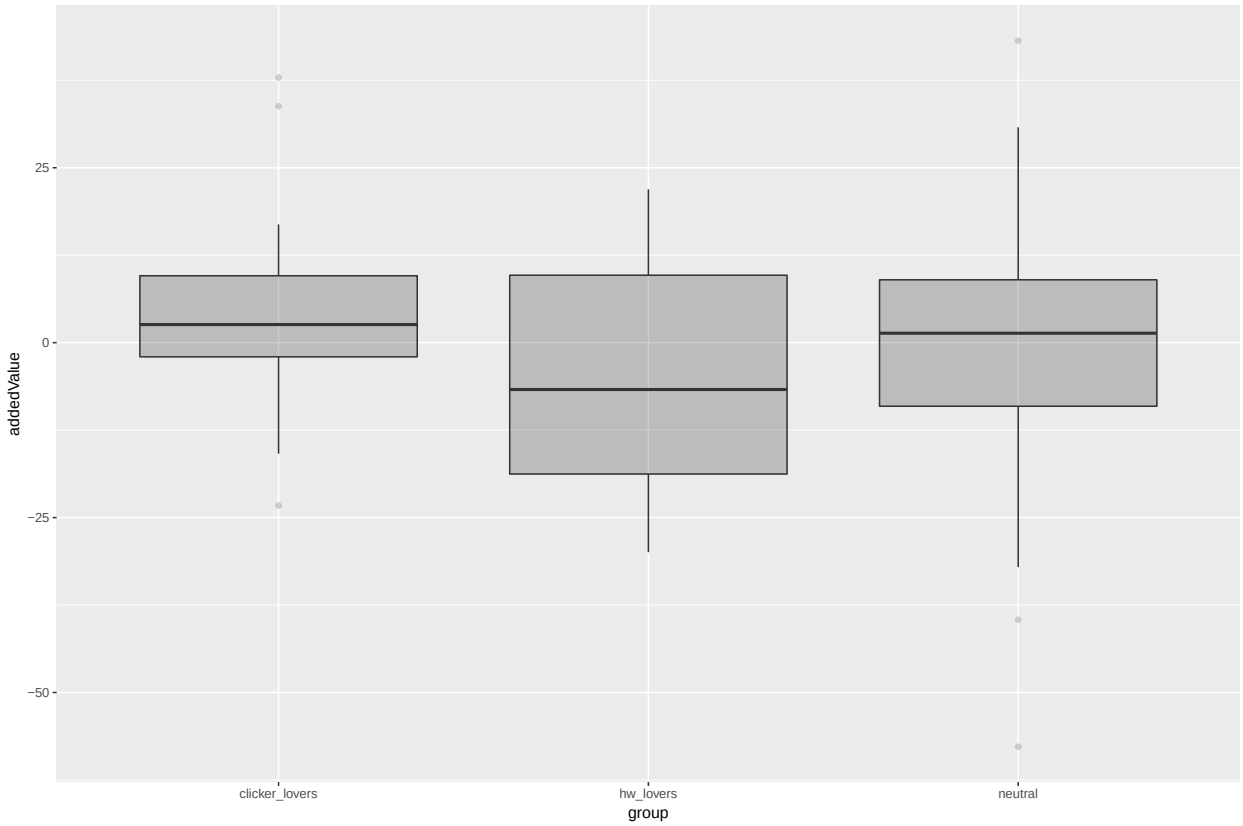
- They are not equal
- Added values of clicker-lovers are higher than those of HW-lovers

There are different ways to identify clicker lovers and HW-lovers.

Identifying HW-lovers and clicker-lovers based on the ratio y/x

Recall that x is the scaled HW score and y is the scaled clicker score. The ratio y/x measures preference towards clickers or homework. For instance, $y/x \approx 1$ for students who invested equal amount of effort into HW and clickers, it is close to 0 for students who invested more effort into HW than into clickers, and it is a large or even infinite positive number for students who invested more effort into clickers than HW.

We choose a threshold q and define group 1 (HW-lovers) to be students with y/x below the q-quantile and group 2 (clicker-lovers) to be students with y/x above the $(1-q)$-quantile. All students who are not in groups 1 or 2 can be thought as "neutral" towards HW or clickers. Let's begin with setting $q = 0.1$.



The number of HW-lovers is 21; the number of clicker-lovers is 21.

Now let's do statistical tests:

- Mann-Whitney-Wilcoxon u-test with the null hypothesis that the true median added value in the group of HW-lovers is the same as the true median added value in the group of clicker-lovers; two-sided and one-sided.
- Welch t-test with the null hypothesis that the true mean added value in the group of HW-lovers is the same as the true mean added value in the group of clicker-lovers; two-sided and one-sided.

```
##
##  Wilcoxon rank sum test
##
## data:  X$addedValue[X$group == "hw_lovers"] and X$addedValue[X$group == "clicker_lovers"]
## W = 169, p-value = 0.2018
## alternative hypothesis: true location shift is not equal to 0
```

```
##
```

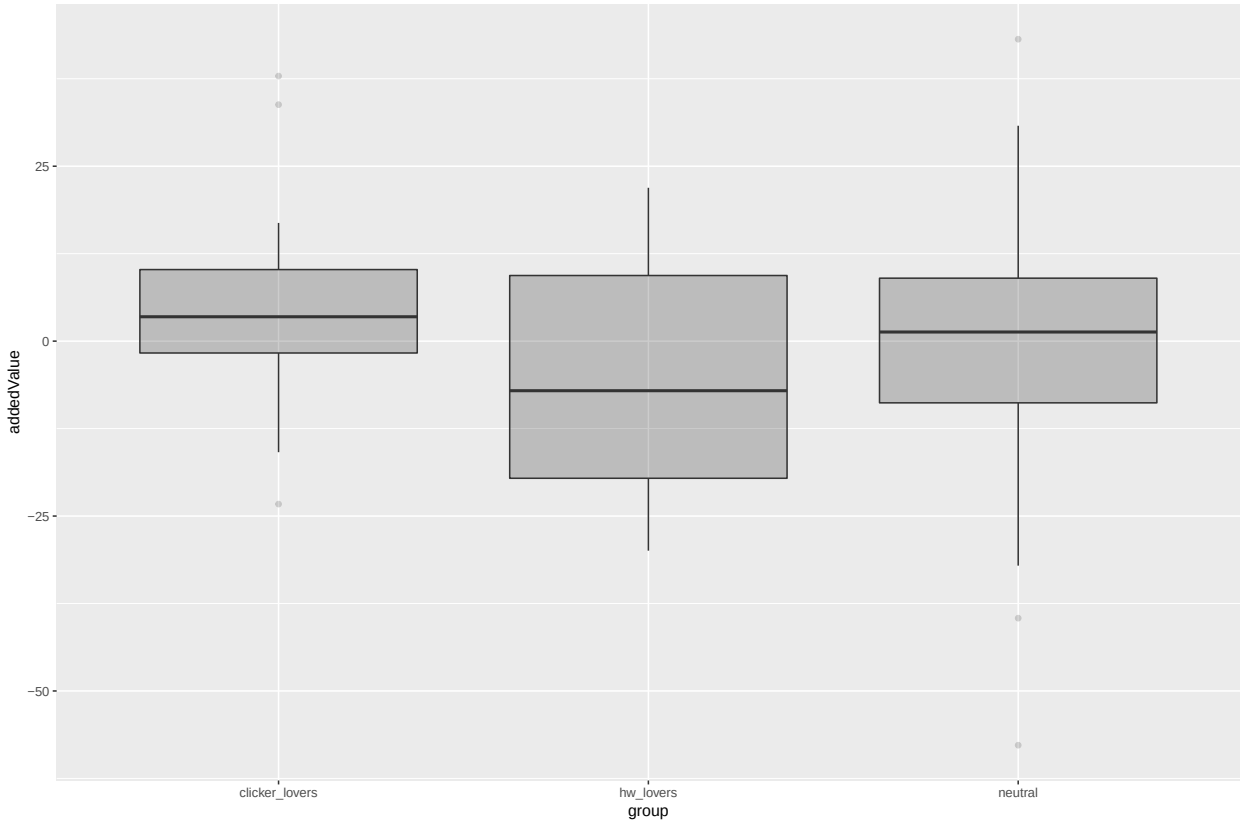
```
##  Wilcoxon rank sum test
##
## data:  X$addedValue[X$group == "hw_lovers"] and X$addedValue[X$group == "clicker_lovers"]
## W = 169, p-value = 0.1009
## alternative hypothesis: true location shift is less than 0
```

```
##
##  Welch Two Sample t-test
##
## data:  X$addedValue[X$group == "hw_lovers"] and X$addedValue[X$group == "clicker_lovers"]
## t = -1.7673, df = 39.102, p-value = 0.08499
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -17.925976   1.207351
## sample estimates:
## mean of x mean of y
## -3.758252  4.601060
```

```
##
##  Welch Two Sample t-test
##
## data:  X$addedValue[X$group == "hw_lovers"] and X$addedValue[X$group == "clicker_lovers"]
## t = -1.7673, df = 39.102, p-value = 0.04249
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.3902481
## sample estimates:
## mean of x mean of y
## -3.758252  4.601060
```

Note that results depend on the type of a test for statistical significance that we use. For instance, the common one-sided t-test gives us the p-value that sufficient to establish statistical significance at the level $p = 0.1$. However, it may not be appropriateto use the t-test here since the data do not follow the normal distribution. The non-parametric Mann-Whitney-Wilcoxon u-test is probably more appropriate.

However, if we change the threshold from $q = 0.1$ to $q = 0.09$, i.e., define the groups of HW-lovers and clicker-lovers to be top and bottom 9% of the metric y/x instead of top and bottom 10%, then the difference in both median and mean between HW-lovers and clicker-lovers become statistically significant:




```
##
##  Wilcoxon rank sum test
##
## data:  X$addedValue[X$group == "hw_lovers"] and X$addedValue[X$group == "clicker_lovers"]
## W = 137, p-value = 0.07075
## alternative hypothesis: true location shift is less than 0
```

```
##
##  Welch Two Sample t-test
##
## data:  X$addedValue[X$group == "hw_lovers"] and X$addedValue[X$group == "clicker_lovers"]
## t = -1.9527, df = 35.549, p-value = 0.02938
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -1.322371
## sample estimates:
## mean of x mean of y
## -4.845730  4.941164
```

Let us now experiment with different values of thresholds. For threshold values q between 0.05 and 0.2, we identify the group of HW-lovers and the clicker-lovers depending on q and report q, N (the sample size, which is equal for the two groups), median and mean in group 1 (HW-lovers) and group 2 (clicker-lovers), Cohen's d (effect size), and p-values of the one-sided u-test (for statistical significance of the difference in samle medians) and the t-test (for statistical siginificance of the difference in means):

```
##       q  N   med.gr.1 med.gr.2 mean.gr.1 mean.gr.2         d   u.test.p
## 1  0.05 11 -13.572168 5.925253 -5.996278  4.680585 0.6533533 0.11755831
## 2  0.06 13 -13.572168 4.574476 -7.526629  4.672423 0.7836651 0.06269946
## 3  0.07 15 -11.459540 2.587630 -4.128905  3.930037 0.5086982 0.18343760
```

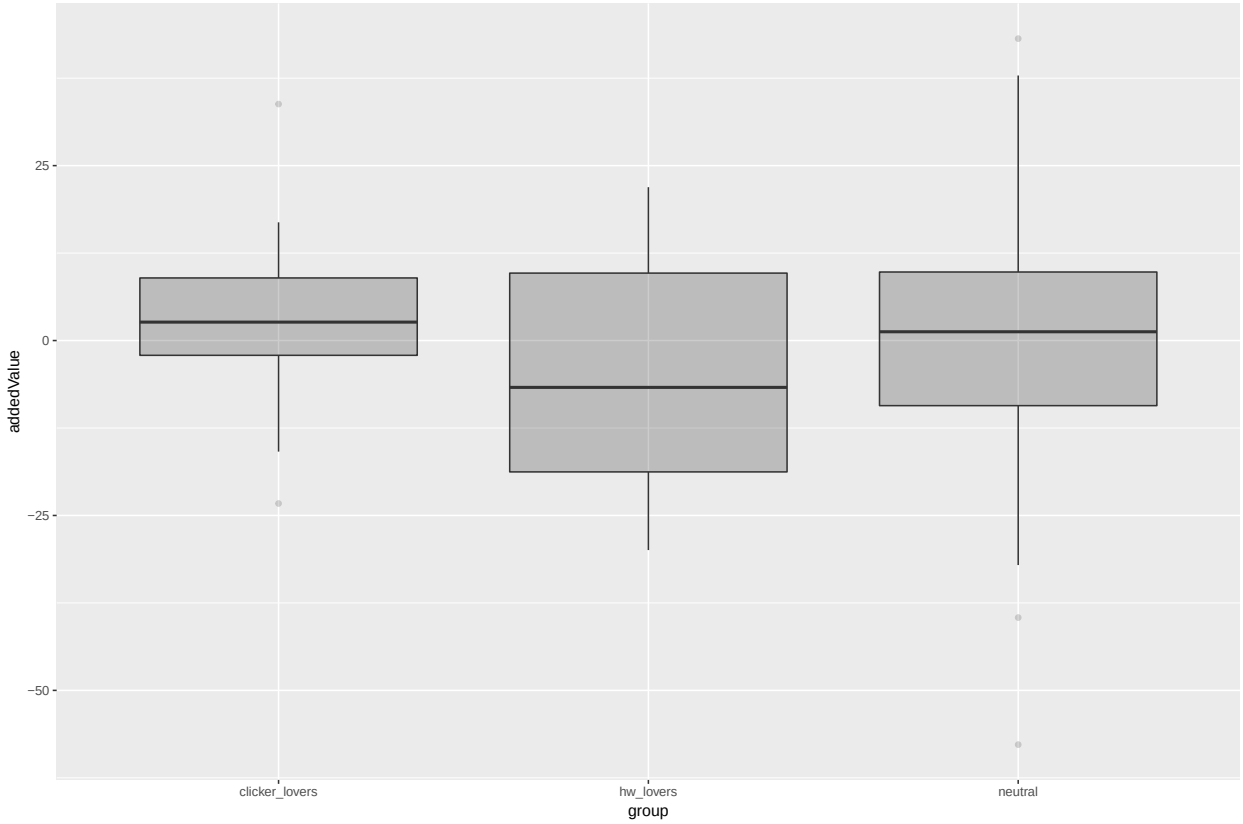
```
## 4  0.08 17 -11.459540 2.587630 -4.605212  3.889213 0.5489983 0.16113982
## 5  0.09 19  -7.084667 3.484537 -4.845730  4.941164 0.6280487 0.07074760
## 6  0.10 21  -6.695595 2.587630 -3.758252  4.601060 0.5453921 0.10088323
## 7  0.11 23  -6.695595 2.587630 -2.861639  4.247947 0.4677318 0.12344076
## 8  0.12 26  -4.055691 2.633812 -2.943035  4.358716 0.4827230 0.13646305
## 9  0.13 27  -6.695595 2.633812 -3.360299  4.226790 0.5136365 0.10740257
## 10 0.14 29  -6.695595 2.587630 -3.187259  2.511103 0.3680961 0.16491134
## 11 0.15 31  -6.695595 2.587630 -3.085079  2.287499 0.3599894 0.15159176
## 12 0.16 33  -6.695595 2.587630 -3.237143  2.287499 0.3741865 0.12249629
## 13 0.17 35  -6.695595 1.867207 -3.644038  2.140916 0.4008498 0.08157149
## 14 0.18 37  -6.695595 2.446251 -3.437457  2.207800 0.3984325 0.08537416
## 15 0.19 39  -6.695595 1.867207 -3.744409  1.338621 0.3604542 0.08754546
## 16 0.20 41  -5.543155 2.446251 -3.414600  1.737496 0.3718879 0.07834230
## 17 0.21 43  -3.353239 1.867207 -2.970208  1.421706 0.3201844 0.10579864
## 18 0.22 45  -3.353239 1.609910 -2.824076  1.070764 0.2852761 0.12967817
## 19 0.23 47  -2.581640 1.867207 -2.140089  1.417224 0.2590083 0.16734727
## 20 0.24 49  -3.353239 1.609910 -2.305030  1.504901 0.2840249 0.11914846
## 21 0.25 52  -3.475317 1.867207 -2.082540  2.028098 0.3060828 0.08263348
## 22 0.26 54  -3.965561 2.156729 -3.246946  2.072294 0.3721095 0.05012502
## 23 0.27 56  -3.965561 2.156729 -3.505449  1.892746 0.3766911 0.05192520
## 24 0.28 58  -3.475317 1.609910 -2.891303  1.225236 0.2834935 0.13057909
## 25 0.29 60  -2.967440 1.609910 -2.597256  1.228484 0.2662521 0.15366033
## 26 0.30 62  -2.015851 2.156729 -2.272186  1.581966 0.2702506 0.15221841
##      t.test.p
## 1  0.06772723
## 2  0.02866927
## 3  0.08762535
## 4  0.06013310
## 5  0.02938158
## 6  0.04249343
## 7  0.06005936
## 8  0.04494107
## 9  0.03255868
## 10 0.08329461
## 11 0.07846372
## 12 0.06675066
## 13 0.04910962
## 14 0.04547869
## 15 0.05781809
## 16 0.04807577
## 17 0.07071079
## 18 0.08879844
## 19 0.10623747
## 20 0.08068803
## 21 0.06008627
## 22 0.02796707
## 23 0.02440945
## 24 0.06485748
## 25 0.07375446
## 26 0.06754019
```

The fact that mean and median added values in group 2 are higher than in group 1 persists regardless of the sample sizes. However, this result is only statistically significant for certain values of sample size.

Identifying HW-lovers and clicker-lovers based on the difference $y - x$

Recall that x is the scaled HW score and y is the scaled clicker score. The difference $y - x$ measures preference towards clickers or homework too. For instance, $y - x \approx 0$ for students who invested equal amount of effort into HW and clickers, it is negative for students who invested more effort into HW than into clickers, and positive for students who invested more effort into clickers than HW.

We choose a threshold q and define group 1 (HW-lovers) to be students with $y - x$ below the q-quantile and group 2 (clicker-lovers) to be students with $y - x$ above the $(1 - q)$-quantile. All students who are not in groups 1 or 2 can be thought as "neutral" towards HW or clickers. Let's begin with setting $q = 0.1$. Below is a whisker chart of three groups:



Now let's conduct statistical tests to see if the findings that median and mean added values for clicker-lovers is higher than for HW-lovers depending on the threshold value.

```
##       q  N   med.gr.1 med.gr.2 mean.gr.1 mean.gr.2         d   u.test.p
## 1  0.05 16 -12.515854 5.925253 -5.463050  2.586452 0.4913468 0.18151433
## 2  0.06 17 -11.459540 5.925253 -4.605212  4.680585 0.5534690 0.12246294
## 3  0.07 17 -11.459540 4.574476 -4.605212  4.672423 0.5631295 0.13143158
## 4  0.08 20  -6.890131 2.587630 -3.875375  3.930037 0.5006252 0.12689053
## 5  0.09 21  -6.695595 2.587630 -3.758252  3.930037 0.5003492 0.11924263
## 6  0.10 22  -6.890131 2.587630 -3.960469  3.930037 0.5203943 0.09555217
## 7  0.11 22  -6.890131 2.587630 -3.960469  3.889213 0.5320862 0.09506741
## 8  0.12 24  -6.890131 2.587630 -3.968353  4.601060 0.5431705 0.08516466
## 9  0.13 28  -6.890131 2.587630 -3.621210  4.247947 0.5167394 0.08570206
## 10 0.14 30  -4.055691 2.587630 -2.832036  4.247947 0.4689212 0.12469710
## 11 0.15 31  -6.695595 2.633812 -3.085079  4.358716 0.5055511 0.09168055
## 12 0.16 33  -6.695595 2.633812 -3.237143  4.226790 0.5219317 0.07523840
```

```
## 13 0.17 34  -5.024417 2.610721 -3.240558  2.768094 0.4029095 0.09098434
## 14 0.18 36  -6.890131 2.227419 -3.854067  2.164102 0.4162032 0.06284752
## 15 0.19 36  -6.890131 1.867207 -3.854067  2.140916 0.4176277 0.06260643
## 16 0.20 38  -6.119375 1.867207 -3.492870  2.140916 0.3948563 0.07208770
## 17 0.21 41  -5.543155 1.867207 -3.414600  1.338621 0.3404679 0.10254108
## 18 0.22 42  -4.448197 2.156729 -3.394768  1.455327 0.3515792 0.08460194
## 19 0.23 43  -3.353239 2.156729 -2.970208  1.474593 0.3221569 0.10577034
## 20 0.24 48  -2.967440 2.156729 -2.262766  1.474593 0.2715944 0.15780403
## 21 0.25 51  -3.597395 1.609910 -2.384581  1.358765 0.2794204 0.10066781
## 22 0.26 51  -3.597395 1.609910 -2.384581  1.358765 0.2794204 0.10066781
## 23 0.27 51  -3.597395 1.352612 -2.384581  1.301274 0.2780796 0.10294970
## 24 0.28 52  -3.475317 1.867207 -2.082540  2.028098 0.3060828 0.08263348
## 25 0.29 52  -3.475317 1.867207 -2.082540  2.028098 0.3060828 0.08263348
## 26 0.30 56  -3.965561 1.609910 -3.505449  1.225236 0.3271667 0.07948998
##      t.test.p
## 1  0.10216338
## 2  0.07188339
## 3  0.06393227
## 4  0.07019189
## 5  0.06810686
## 6  0.05915203
## 7  0.04905287
## 8  0.03624852
## 9  0.03380453
## 10 0.04499800
## 11 0.03011681
## 12 0.02229987
## 13 0.05582669
## 14 0.04276152
## 15 0.04126688
## 16 0.04753364
## 17 0.06564544
## 18 0.05740619
## 19 0.07047078
## 20 0.09981052
## 21 0.08349229
## 22 0.08349229
## 23 0.08168635
## 24 0.06008627
## 25 0.06008627
## 26 0.04231656
```

Now statistical significance is achieved when the threshold is 15% or 16%.

Identifying HW-lovers and clicker-lovers based on the x and y separately.

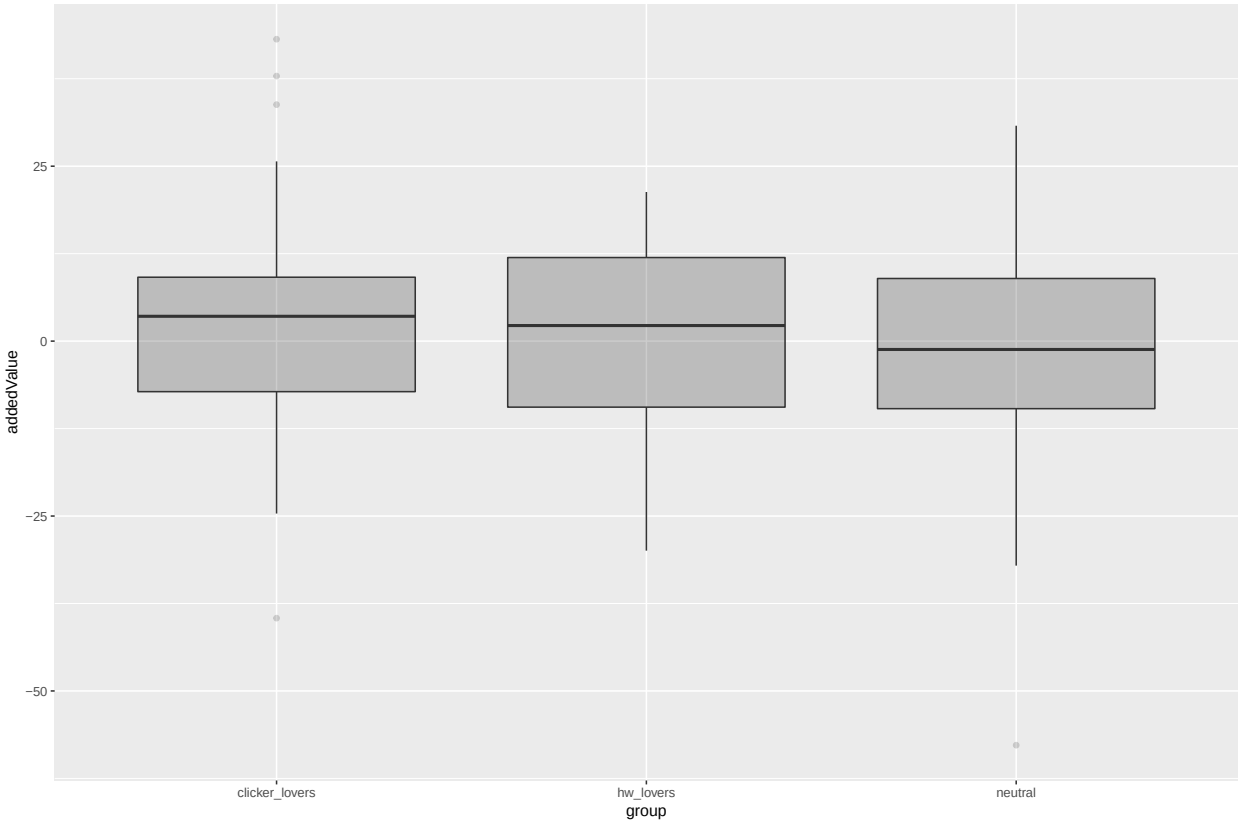
Recall that x is the scaled HW score and y is the scaled clicker-score. Let us say that HW-lovers are students whose HW score is above the median while clicker score is below the median. Likewise, clicker-lovers are students whose clicker score is above the median while HW score is below the median. The rest are "neutral".

Below are some statistics:

```
##            group   N    sigma    median       mean
## 1 clicker_lovers  52 14.81492  3.549314  2.7120354
## 2      hw_lovers  51 12.75968  2.225535  0.3745431
```

```
## 3        neutral 102 15.04744 -1.192545 -1.5698778
```

Here is the whisker chart



Now let's perform statistical tests:

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  X$addedValue[X$group == "hw_lovers"] and X$addedValue[X$group == "clicker_lovers"]
## W = 1254, p-value = 0.6372
## alternative hypothesis: true location shift is not equal to 0
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  X$addedValue[X$group == "hw_lovers"] and X$addedValue[X$group == "clicker_lovers"]
## W = 1254, p-value = 0.3186
## alternative hypothesis: true location shift is less than 0
```

```
##
##  Welch Two Sample t-test
##
## data:  X$addedValue[X$group == "hw_lovers"] and X$addedValue[X$group == "clicker_lovers"]
## t = -0.85852, df = 99.351, p-value = 0.3927
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.739701  3.064716
## sample estimates:
## mean of x mean of y
```

```
## 0.3745431 2.7120354
```

```
##
##  Welch Two Sample t-test
##
## data:  X$addedValue[X$group == "hw_lovers"] and X$addedValue[X$group == "clicker_lovers"]
## t = -0.85852, df = 99.351, p-value = 0.1963
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 2.183118
## sample estimates:
## mean of x mean of y
## 0.3745431 2.7120354
```

We could go further, set several thresholds and define HW-lovers to be students whose HW score is above a certain threshold and whose clicker score is below a certain, second, threshold; and clicker-lovers to be students whose clicker score is above a third threshold and HW score is below a fourth threshold. We could become even more creative and invent some other way to define HW-lovers and clicker-lovers. This demonstrates that data can be manipulated with to produce more impressive results to increase statistical significance.

Regression analysis

An alternative approach to show that clickers were a more effective teaching method in the course than handwritten homework is conducting standard regression analysis. The response variable is the added value and regressors are HW, clicker, or both HW and clicker scores.

We find that there is no significant correlation between added values and HW score, but there is a positive correlation between added values and clicker score, i.e., students with higher clicker scores tend to achieve higher results in the course. This positive correlation is statistically significant at the level $p = 0.1$.

Regression of added values vs HW score

```
##
## Call:
## lm(formula = X$addedValue ~ X$writtenHW)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.826  -9.313   1.224   9.339  43.120
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.11590    4.85056   0.230    0.818
## X$writtenHW -0.02260    0.09605  -0.235    0.814
##
## Residual standard error: 14.53 on 203 degrees of freedom
## Multiple R-squared:  0.0002726,  Adjusted R-squared:  -0.004652
## F-statistic: 0.05535 on 1 and 203 DF,  p-value: 0.8142
```

This shows that an increase by 1% in HW score is linked to a decrease by 0.006 in added value. The fact that the sign of the regression coefficient is negative is not statistically significant.

Regression of added values vs clicker score:

```
##
## Call:
## lm(formula = X$addedValue ~ X$clickers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.302  -9.912   0.754  10.453  42.533
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.15183    3.61151  -1.703   0.0900 .
## X$clickers   0.11474    0.06469   1.774   0.0776 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.42 on 203 degrees of freedom
## Multiple R-squared:  0.01526,    Adjusted R-squared:  0.01041
## F-statistic: 3.146 on 1 and 203 DF,  p-value: 0.07761
```

This shows that an increase by 1% in HW score is linked to a decrease by 0.111 in added value, i.e., there is a small positive relationship. The fact that the sign of the regression coefficient is positive is statistically significant at the level $p = 0.1$.

Two-variable regression

```
##
## Call:
## lm(formula = X$addedValue ~ X$writtenHW + X$clickers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.405  -9.087   0.702   9.940  42.449
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.12239    5.61935  -0.734   0.4640
## X$writtenHW -0.04547    0.09633  -0.472   0.6374
## X$clickers   0.11878    0.06537   1.817   0.0707 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.44 on 202 degrees of freedom
## Multiple R-squared:  0.01635,    Adjusted R-squared:  0.006607
## F-statistic: 1.678 on 2 and 202 DF,  p-value: 0.1893
```

When both HW score and clicker score are used as regressors, the negative relation between added value and HW score becomes stronger - now an increase by 1% in HW score is linked to a decrease by 0.028 in added value. However, negativity of the relation between added value and the HW score is still not statistically significant.