Analyzing FCI Data from Students after Historical Discussions

Purpose

Physics Education Research (PER) has sought to improved learning of physics concepts, and since the 1970s it has been noted that students have intuitions about the world similar to medieval concepts. The question becomes: would having the students understand how their ideas fit into the history of science actually help them understand the science better and improve their conceptual framework as measured by their adherence to Newtonian mechanics?

Preliminiary discussion

The standard tool for measuring student concepts is the **Force Concept Inventory** or simply the **FCI**. This is an inventory of 30 questions that queries students and sees if they provide either the correction (Newtonian) answer or instead give one of the other four distractor answers. Students do not randomly pick wrong answers, but instead they cluster towards particular distractors for most questions. Those popular distractors have much in common with ideas from pre-Newtonian mechanics, especially the medieval concept of *impetus*. The researcher will note how students change towards the Newtonian view by having given the FCI test at the beginning of a physics course (the pretest), and then again towards the end of the course (the posttest); sometimes a delayed posttest is also administered. Commonly PER researchers have used a measure of learning called the normalized gain, g. Taking one's pretest score (out of 100%) and ones posttest score, and the normalized gain figures how much of the possible learning (the difference between prescore and 100%) took place.

$$g = \frac{post - pre}{100 - pre}$$

For a class of students, one may either provide the average postscore and average prescore of the class (brackets indicate taking the average),

$$< g > \quad = \quad \frac{< post > - < pre >}{100 - < pre >}$$

but preference is given here to taking the average of each student's noramlized gains. This will better capture the spread of individual students. Moreover, it forces the calculation of g to be based on matched student data–that is, individual students having both a pre- and a posttest score.

$$g_{ave} \quad = \quad < \frac{post - pre}{100 - pre} >$$

For larger classes this difference may not matter, but the sample size for this study is small (less than 100 students total, less than 30 per class). Additionally, the standard measure of how much an effect the educator was on their students' understanding is the effect size, as measured in the following way:

$$d = \frac{< post > - < pre >}{\sigma}$$

Here, σ represents the standard deviation in scores, and d is the effect size. Generally, $d\ 0.2$ is considered small, and $d > 0.8$ is a significant effect.

Study Design

In this study, the question to be considered is: how does teaching students about the history of physics help them understand the physical concepts themselves? The hypothesis is that thinking and discussing how scienists came to the Newtonian view of the world will help them understand it better, especially comparing it to the impetus view it replaced. This is considered likely because students already hold impetus-like intuitions, so having the history of science confront their gut feelings may help them see how the Newtonian conception works better.

To test this, students were in two groups: those who would have the normal class lectures, homework, labs, and tests; and those who would have additional discussion and projects related to the history of physics. This included readings from the works of Galileo (translated and explained), at-home experiments, and drawing free-body diagrams. The experimental group had two such projects, one in each of the first two quarters of the school year.

All students were in first-year high school physical science courses, and all students had the same teacher (me). The pretest FCI scores between the two groups were notably different The experiemtnal group's mean FCI score was 9.12, and the control group's mean was 6.79. The two groups were within the standard deviations of their respective populations (3.37 and 2.85 respectively). A t-test statistic was applied and found that the difference between the groups was statistically significant $(t = 2.82,\ df = 72,\ p < 0.01)$. This means that the populations were different, and that difference needs to be accounted for in the overall analysis.

Let us now consider how much learning took place. First, we consider the normalized gains:

```
## [1] Experimental Normalized Gain: 0.258110891753516
```

```
## [1] Control Normalized Gain: 0.092730918014465
```

```
## [1] Exp/Cont Gains Ratio: 2.78343941028659
```

The experimental group clearly out-performed, having triple the normalized gains over the control group. However, a gain of 26% in PER is considered small, and the control group's gains are very weak. For completeness, we need to also look at effect sizes.

```
## [1] Experimental Effect Size: 1.23729889079386
```

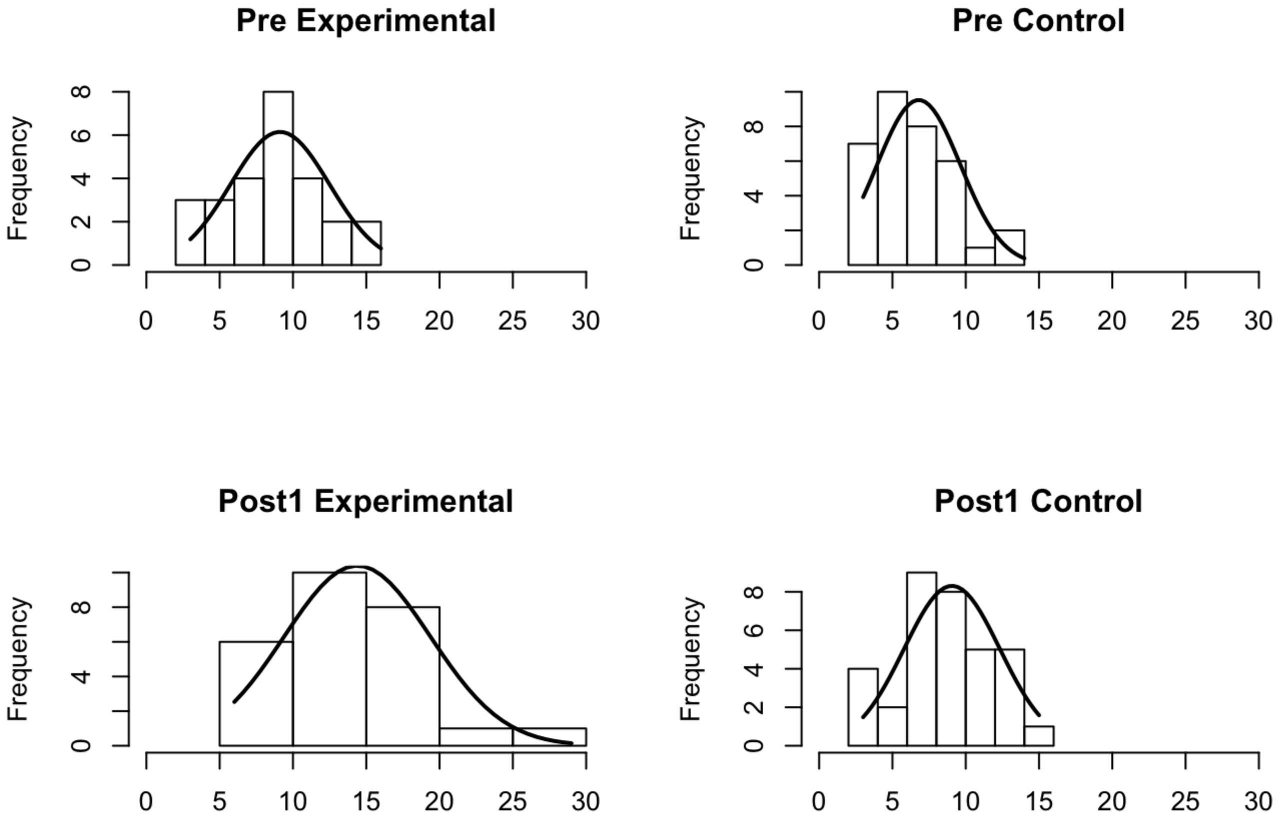
```
## [1] Control Effect Size: 0.739237075685193
```

```
## [1] Exp/Cont Effect Size Ratio: 1.67375113003771
```

Again, the experimental group performs better, and the effect size is nearing twice that of the control group. In both cases, the effect size is very large.

Group Plots

The distribution of FCI scores at the beginning of the school year and the first posttest are provided here, comparing the experimental and control groups.



A curve representing the normal distribution is added, showing that in all cases the scores are following a single-peaked normal curve, though there is noticable skew in a few cases. The change in the control group is small, especially compared to the experimental group.

Controlling for Confounding Factors

While these results indicate a very strong effect from the educational method change, other factors are measured to consider if another variable is at play. Two piece of demographic information were gathered from students: racial background and education level of parents/guardians. Additionally, students were administered the Lawson Scientific Reasoning test to see if learning gains could be accounted for by some students already being good at scientific reasoning.

Analysis of the pretest scores found that only one variable consistently correlated with these initial scores: the Lawson test score. Educational background of the parent or guardian had no significant effect, and only the racial category of Hispanic/Latinx met statistical significance, and it was a positive effect (about 4 more correct reponses to the FCI pretest than white students); other racial categories were not significant and were negative. However, even this racial effect could be by chance due to the multiple racial category comparisons. Stricker rules on statistical significance would suggest that racial categories had no appreciable affect on FCI prescores.

```
##
## Call:
## lm(formula = pre_result ~ lawson_result + Period.x + Race + Highest.level.of.educatio
n.of.parent.guardian,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5693 -1.6110  0.0778  1.3918  7.3517
##
## Coefficients:
##                                                                                 Estima
te
## (Intercept)                                                                       3.44
54
## lawson_result                                                                     0.48
47
## Period.xG                                                                        -0.41
61
## RaceBlack/African American                                                       -0.74
67
## RaceHispanic/Latinx                                                               4.36
63
## RaceAsian                                                                        -0.41
94
## Highest.level.of.education.of.parent.guardianSome college                        -2.87
53
## Highest.level.of.education.of.parent.guardianBachelor's (4-year) degree         -2.06
40
## Highest.level.of.education.of.parent.guardianGraduate school--no degree earned  -4.15
43
## Highest.level.of.education.of.parent.guardianGraduate school--Masters           -2.13
65

## Highest.level.of.education.of.parent.guardianGraduate school--PhD               -1.27
76
##                                                                                 Std. E
rror
## (Intercept)                                                                         1.
8738
## lawson_result                                                                       0.
1239
## Period.xG                                                                           1.
1014
## RaceBlack/African American                                                          2.
5899
## RaceHispanic/Latinx                                                                 2.
0191
## RaceAsian                                                                           2.
1408
## Highest.level.of.education.of.parent.guardianSome college                           3.
2210
## Highest.level.of.education.of.parent.guardianBachelor's (4-year) degree             2.
1087
```

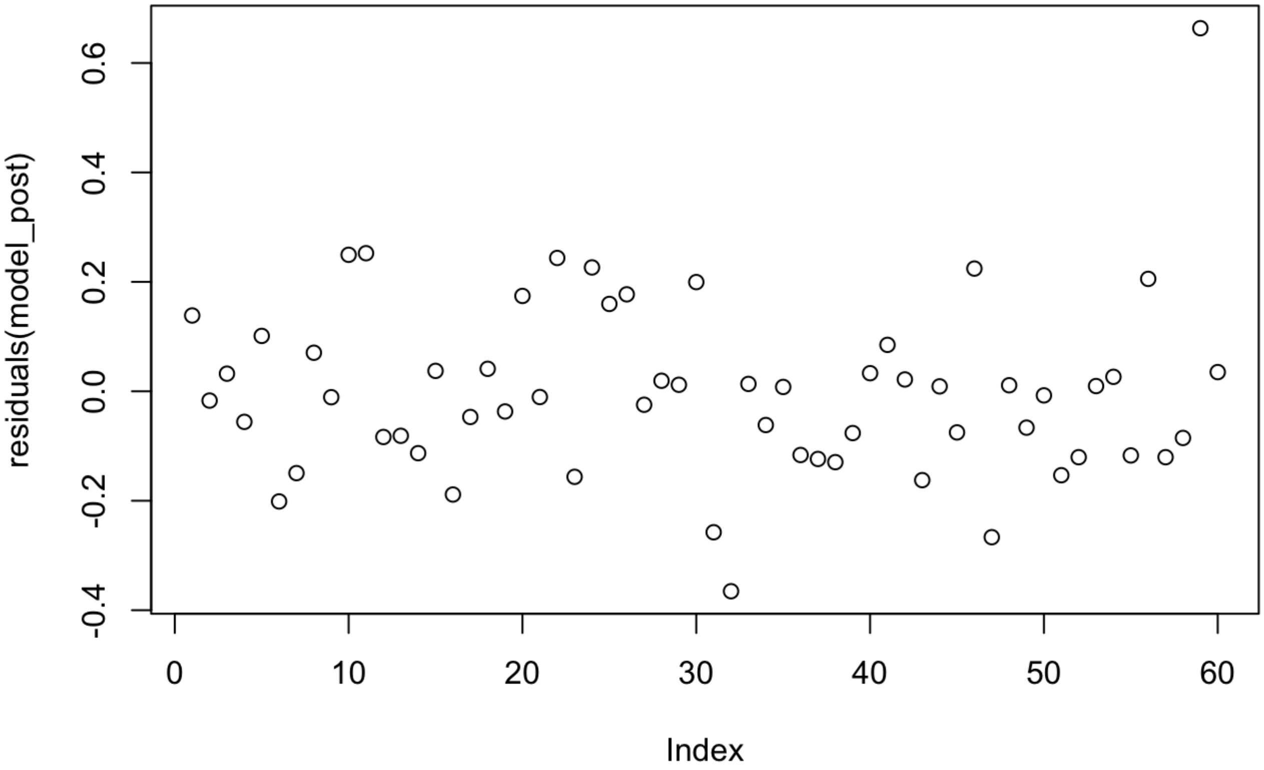
```
## Highest.level.of.education.of.parent.guardianGraduate school--no degree earned     2.
1858
## Highest.level.of.education.of.parent.guardianGraduate school--Masters              1.
8991
## Highest.level.of.education.of.parent.guardianGraduate school--PhD                  2.
0818
##                                                                                t valu
e
## (Intercept)                                                                      1.83
9
## lawson_result                                                                    3.91
3
## Period.xG                                                                       -0.37
8
## RaceBlack/African American                                                      -0.28
8
## RaceHispanic/Latinx                                                              2.16
3
## RaceAsian                                                                       -0.19
6
## Highest.level.of.education.of.parent.guardianSome college                       -0.89
3
## Highest.level.of.education.of.parent.guardianBachelor's (4-year) degree         -0.97
9
## Highest.level.of.education.of.parent.guardianGraduate school--no degree earned  -1.90
1
## Highest.level.of.education.of.parent.guardianGraduate school--Masters           -1.12
5
## Highest.level.of.education.of.parent.guardianGraduate school--PhD               -0.61
4
##                                                                                Pr(>|t
|)
## (Intercept)                                                                    0.0732
00

## lawson_result                                                                  0.0003
36
## Period.xG                                                                      0.7075
12
## RaceBlack/African American                                                     0.7745
75
## RaceHispanic/Latinx                                                            0.0364
65
## RaceAsian                                                                      0.8456
47
## Highest.level.of.education.of.parent.guardianSome college                      0.3772
39
## Highest.level.of.education.of.parent.guardianBachelor's (4-year) degree        0.3334
17
## Highest.level.of.education.of.parent.guardianGraduate school--no degree earned 0.0643
97
## Highest.level.of.education.of.parent.guardianGraduate school--Masters          0.2671
27
## Highest.level.of.education.of.parent.guardianGraduate school--PhD              0.5427
87
```

```
##                                                                                    
## (Intercept)                                                                     .  
## lawson_result                                                                   ***
## Period.xG                                                                          
## RaceBlack/African American                                                         
## RaceHispanic/Latinx                                                             *  
## RaceAsian                                                                          
## Highest.level.of.education.of.parent.guardianSome college                          
## Highest.level.of.education.of.parent.guardianBachelor's (4-year) degree            
## Highest.level.of.education.of.parent.guardianGraduate school--no degree earned .  
## Highest.level.of.education.of.parent.guardianGraduate school--Masters              
## Highest.level.of.education.of.parent.guardianGraduate school--PhD                  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.839 on 41 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.408,  Adjusted R-squared:  0.2636 
## F-statistic: 2.826 on 10 and 41 DF,  p-value: 0.009193
```

As for the Lawson test score, this was highly predictive of pretest scores, with a coefficient of 0.375, p << 0.01, and this variable on its own accounted for nearly 25% of the variance. This means that a student who scores a 100% on the 24 questions of the Lawson test would have about 9 more points on the FCI test than someone who scored a 0.

Would the Lawson test also be predictive of overall gains? This was tested with a linear regression model as well, which is detailed below, filtering down to the only variables founds to be notable.

```
## 
## Call:
## lm(formula = g_ind ~ lawson_result + Period.x, data = df)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.36535 -0.11388 -0.00895  0.04845  0.66356 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)  
## (Intercept)   0.010159   0.073442   0.138   0.8905  
## lawson_result 0.007349   0.006036   1.218   0.2284  
## Period.xG     0.131493   0.051071   2.575   0.0127 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.1644 on 57 degrees of freedom
## Multiple R-squared:  0.2234, Adjusted R-squared:  0.1962 
## F-statistic: 8.199 on 2 and 57 DF,  p-value: 0.000742
```

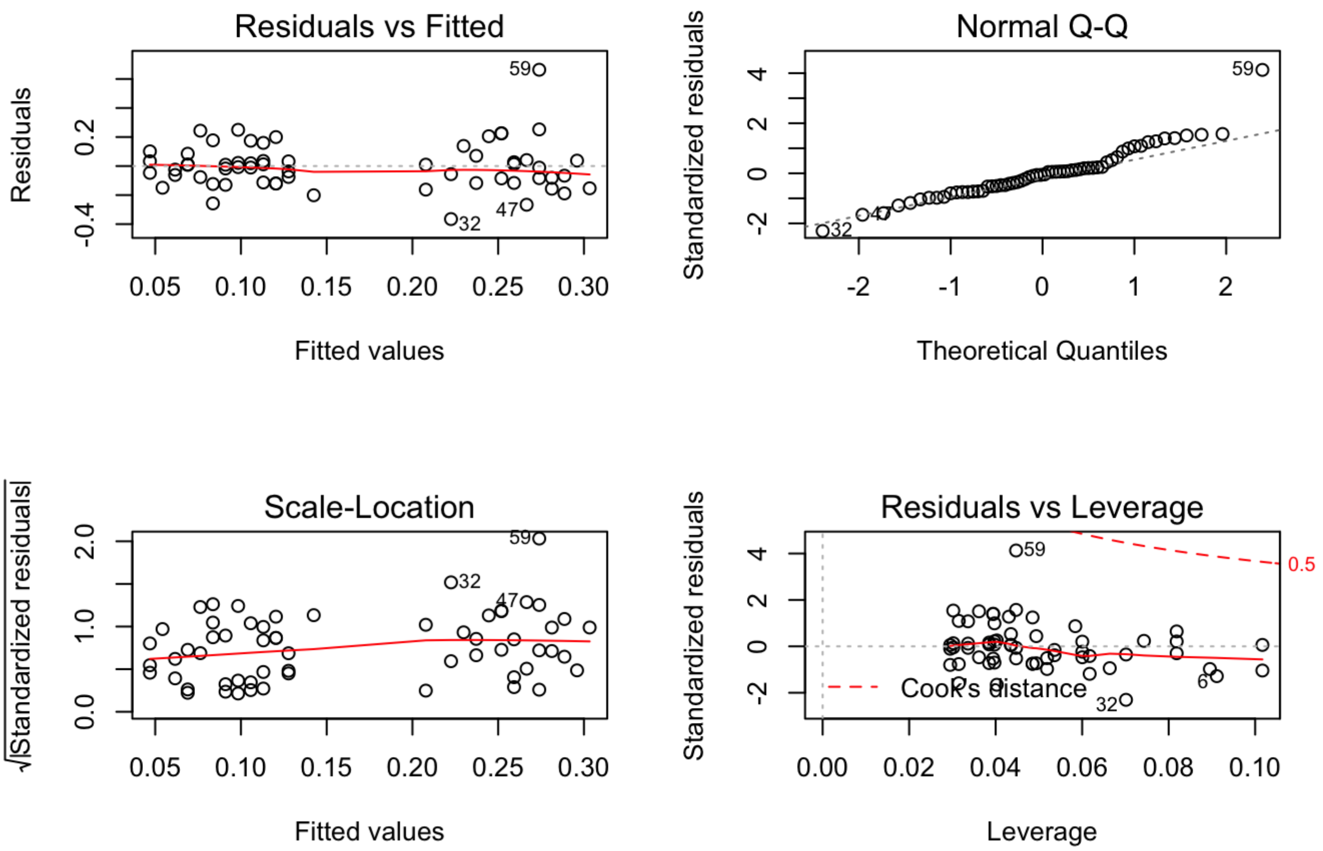


Complex models that included race and/or parential education level gave no significant or consistent resulsts. For example, education level didn’t consistently have an effect–more education sometimes lowers scores, sometimes raises scores. p values were also large, indicating that race and education level were not useful variables in explaining student learning differences.

As for scientific reasoning, on its own it was shown to correlate well with normalized gains, but not nearly as much as which group students were in. As seen in the summary of the general linear model above, the lawson_result (the measure of scientific understanding) had a minimal effect at all (less than 1% increase in normalized gains with each correct answer on the scientific reasoning test; the difference between a 0 and 100% on this test would only explain on average a 17% difference in normalized gains.) Moreover, the effect from scientific reasoning fails to achieve statistical significance $(p > 0.05)$.

Another variable compared to was the pretest scores of the students. This is important considering that the experimental and control groups had significant differences in their average FCI scores. Even as the lone variable, however, pretest score had no predictive value towards the noramlized gains. In other words, students who initially tested well did not have more learning gains that students who initially tested poorly on the FCI.

What remains is the period of instruction. If a student was in G period (the experimental group), and we still try to account for the results along with scientific reasoning ability, they had a 13% larger normalized gain than the control group, and this result passes the significance test $(p < 0.05)$. The model also fits without leaving any patterns in the residues, as see in the plot above. There is one notable outlier, but its removal does not change the results significantly. Further analysis of the residuals also shows nothing that is worrisome, even with the one outlier.



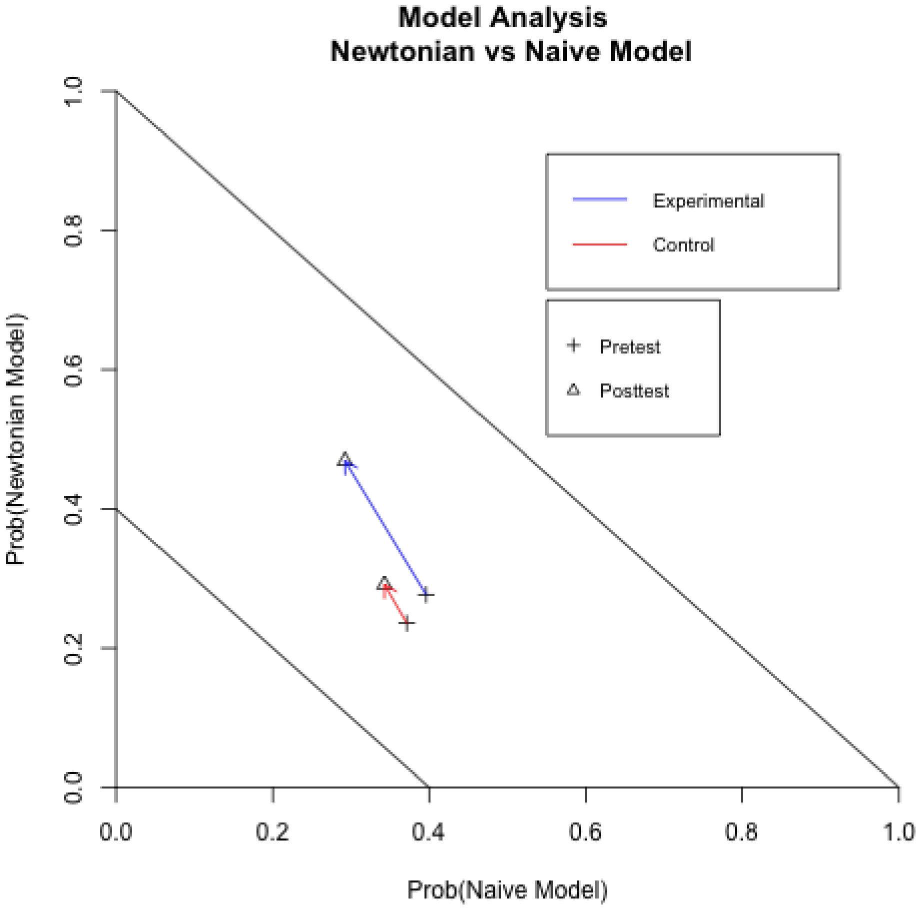
We see no trends in Residuals vs Fitted, Scale-Location, or Residuals vs Leverage, and Q-Q shows something very close to an expected normal distribution of residuals (the residuals are randomly distributed about the mean, not patterned).

However, we do find a small value for R^2, such that only about 22% of the variance in student gains is explained by the model. It is noted that including race or educational level of parents does not improve R^2 by any notable amount, so the remainder of the variance cannot be accounted for with the given data.

Model Analysis

The FCI measures how well students can answer force and motion questions in terms of Newtonian mechanics. However, an FCI score on its own does not tell you what model the students may have in mind. What has been noted is that students tend to come to a physics class with a belief in motion and force similar to the medieval concept of ‘impetus’. Students tend to not randomly choose wrong answers to FCI questions, but instead they tend to cluster towards pre-Newtonian concepts of dynamics. The question becomes: how well has teaching moved students from pre-Newtonian to Newtonian thinking? This is achieved by means of model analysis.

In model analysis, we take the class as a whole and see how often they choose one model versus another. This is turned into a probability of using one model over another, and we can track how these probabilities change over time. This is easier to tell by creating a model analysis graph. The position of a point on the graph will indicate how often they use the Newtonian model, the pre-Newtonian/impetus model, and also how the student answers fail to conform to either model.



Points in the lower-right corner represent students who always use the naive, non-Newtonian model; points in the upper-left corner are students who always use the Newtonian model; point at the origin represent students who always choose something else than the suggested models (naive and Newtonian). A line representing 40% of the two model probabilities combined is added to the graph, helping to show that the two models (naive and Newtonian) together account for the majority of the model space. A line representing 100% is also given, showing the bounds of the model space.

As can be seen, both the experimental group (in blue) and the control group (in red) moves from the more naive end of the model space towards the Newtonian end. We also note that in both cases, the final points move away from the origin, indicating that students chose fewer random answers and more than before conformed to the Newtonian or naive/impetus model. It is also evident that the experimental (blue) group had the largest move towards Newtonian thinking, while the control (red) group had rather little overall conceptual change.