Modelling Assessment Data with a Hierarchical Approach

Abstract

Assessment data were collected from multiple institutions in fall 2013 and winter 2014 to measure the conceptual understanding and perspectives of statistics for students taught under a randomization-based curriculum. We concentrated on identifying student and instructor characteristics that were associated with student gains and determining whether there was a disparity in student gains based on level of instructor's experience with the curriculum. A hierarchical modelling approach was employed to analyze the data, because multiple students were in each section taught by the same instructor and variables were collected on students and instructors. Our exploratory data analysis included graphical analysis and instructor classification based on their experience with the curriculum, teaching and school background, and the idiosyncrasies of their students. Subsequently, we ran hierarchical models to model student improvements in their conceptual understanding of statistics with student-level and instructor-level predictors.

Research Goals

The **primary goals** of our research were:

- To distinguish the characteristics of students and instructors that are linked to student success in an introductory statistics course under a randomization-based curriculum.
- To detect whether there was a disparity in student success based on the instructors' experience with the curriculum.

The **secondary goals** were:

- To understand and apply hierarchical models for hierarchical data.
- To program in *R* for performing analyses and producing relevant graphics.

Background and Significance of Research

Recent educational research found that implementing a randomization-based curriculum in an introductory statistics course can provide a more enriching experience for students compared to learning under a traditional curriculum (Tintle, VanderStoep, Holmes, Quisenberry, and Swanson, 2011). Students are able to focus on being trained to think and reason statistically rather than memorizing formulas to calculate statistics such as standard deviations or correlation coefficients. The benefits of being able to merely memorize a formula are not as impactful as being able to interpret results and relate to practicality (Garfield, 1992).

Reformers in statistics education have been developing randomization-based web applets that helped facilitate the new curriculum (Rossman and Chance 2008). These reformers are deliberate in transforming the traditional statistics classroom into a dynamic learning environment. Currently, there are textbooks developed that encompass the idea of performing randomization tests for statistical inference as opposed to parametric tests (e.g., ISCAM; Chance and Rossman, 2014; Unlocking the Power of Statistics, Lock et al, 2014). Studies have suggested that promoting learning through the use of activities and technology can encourage students to share insights with their peers and exchange feedback with their instructors (Garfield and Ben-Zvi, 2009). These elements are important to students' success because students have the opportunities to construct their own knowledge during the process of learning statistics, e.g., through the web applets, instead of being presented the theoretical results. In contrast, the previously described elements do not appear to exist in a traditional statistics classroom with the instructor lecturing each session rather than facilitating students' learning.

In assessing the effectiveness of a randomization-based curriculum, students who were instructed through the randomization-based curriculum at Hope College were compared to student performance in prior years with a more traditional curriculum. They were also measured on their retention of statistical concepts after having taken the course. It was evident that students at Hope College who followed a randomization-based curriculum tended to perform better on the topics of data collection, data design, and tests of significance from the CAOS test than students who were under the previous curriculum; a follow-up study suggested these students also retained more statistical knowledge (Tintle, Topliff, Vanderstoep, Holmes & Swanson, 2012). The current study aims to delve deeper into the specific attributes of students and instructors who will prosper the most by using this type of curriculum.

Data Analysis Methods

RStudio version 0.98.507 was the primary statistical software used in running our hierarchical models, producing graphical displays, and cleaning our data. Our entire data analysis was divided into two stages that were exploratory and confirmatory. However, we actively shifted between the two stages as we constantly continued to explore relationships of variables to make sense of the results that we found.

Stage 1: Exploratory data analysis

Graphical displays
During the exploratory data analysis stage, the main *R* packages that we used were **ggplot2** and **gridExtra**. We generated multi-dimensional graphs to observe how variables were associated and interacted with each other in predicting student gains on a concept inventory modeled after the CAOS test. These graphs provided us some insights of the results to expect in the latter data analysis stage. Examples of such graphs were boxplots conditional on a third variable and scatterplots with regression lines imposed and color coded based on a third variable. The *R* codes required to create these graphs were an extension to existing functions, because we aimed to create displays that were intriguing and original while still informational. With the graphs that we generated, we used the package **gridExtra** to properly combine related graphs to a single cohesive graph when we wanted to illustrate the idea of unconditional versus conditional associations. These combined graphs were then saved to jpeg files so we can access them directly in the future for presentations and project write-ups.

Cluster analysis
In the latter part of first stage, we decided to perform cluster analyses to obtain two latent variables that grouped instructors based on the observed student-level variables and instructor-level variables, respectively. These two latent variables were treated as instructor level predictors when we later ran hierarchical models. We then compared the within and between sum of squares for differing number of clusters and settled that the most optimal number of clusters to use was four. The practical characterization of each cluster was based on the means of each cluster for the variables that were used.

Comparison of modelling techniques
In addition, we explored three modelling techniques: hierarchical linear modelling (HLM), unpooled modelling, and no pooling. We generated different graphical displays to illustrate the advantages of an HLM through "partial pooling," "borrowing strength," and "shrinkage."

Stage 2: Confirmatory data analysis

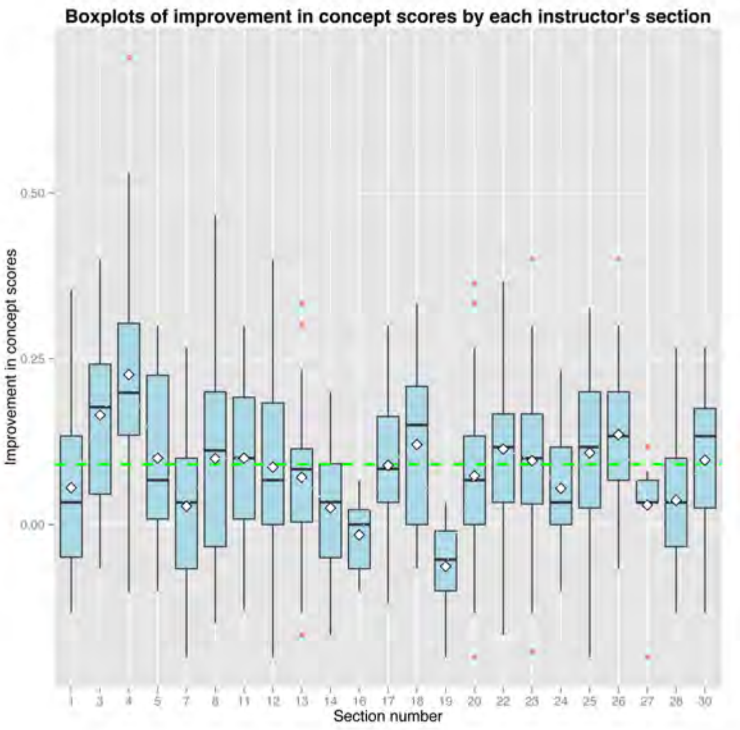
Hierarchical models
When proceeding to the confirmatory data analysis stage, we used the **lme4** and **lmerTest** packages in *R*. The **lme4** package assisted us in fitting hierarchical linear and generalized linear models. We obtained the estimates for the coefficients and variance components after running different models. The first model that we ran was an unconditional means model with improvement in concept scores as the response. With the estimated variance components from this model, we estimated the intraclass correlation coefficient, also called the variance partition coefficient. This estimate gave us a hint as to the importance of defining a hierarchy when performing our analysis. Extending from the unconditional means model, we fitted additional models based on the different combinations of predictors. Our method was to include all predictors in a model and then perform backward elimination to result in a final model. With the

different models that we had, we compared their performance based on AIC values. Graphical displays were also generated to visually communicate the results we discovered from the models we ran.

Exploratory Data Analysis

Example of between-instructor and within-instructor variability in improvements

In our exploratory data analysis stage, we created different graphs to explore the variability of improvement in concepts that is explained by the differences among instructors and the differences within instructors, which is equivalent to the differences among students. Below is an example of a graph that illustrates the above idea. Shown are the boxplots of improvement in concepts separated by sections, with the means plotted in diamonds and outliers plotted in red. We also included a green dotted line that displays the national average improvement of .091 from pre to post on the CAOS test.



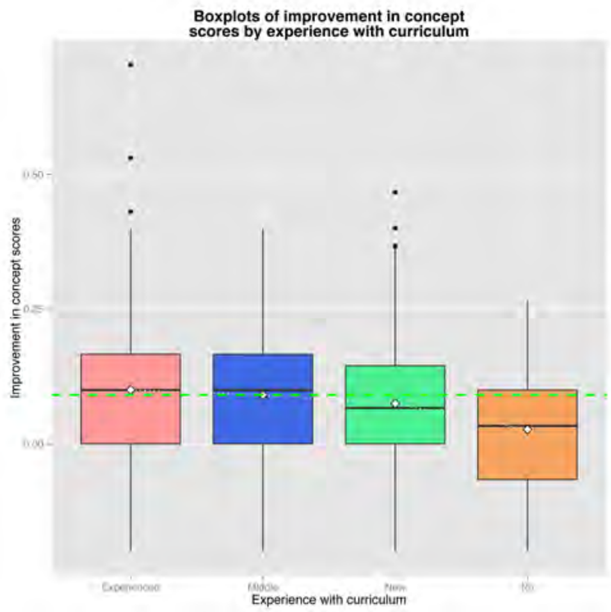
It is evident that some sections had an average improvement that is above the national average, such as sections 3 and 4, and they also had an average improvement that was positive rather than negative. In contrast, some sections had students who performed below the national average and also declined, on average. There was generally more variability within each section than between sections, as most of the boxplots overlapped with each other. However, there were still enough differences among instructors/sections to motivate us to explore potential explanations for these instructor/section level differences.

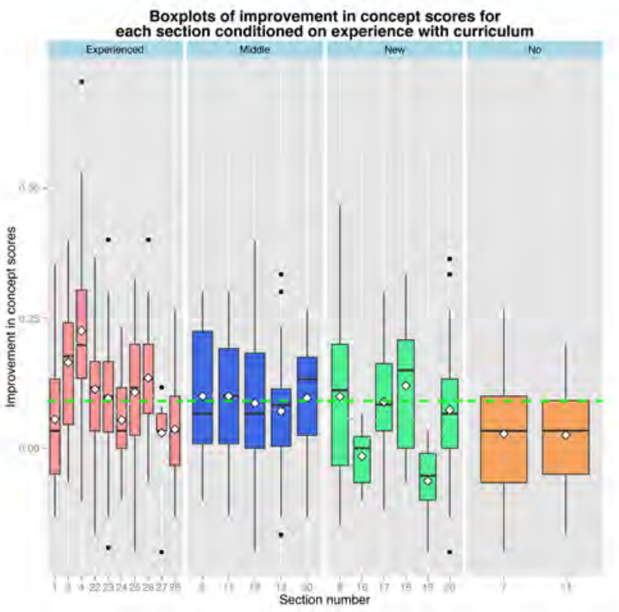
Instructor classification

There were three different approaches that we used to group similar instructors/sections together. The first approach was an ad-hoc method where we classified instructors based on their experience with the curriculum, as we hypothesized that improvements may differ based on this variable. The second and third approaches were based on student characteristics and instructor/section characteristics. For the latter two approaches, we decided to conduct cluster analyses to distinguish instructors by using student level variables and instructor level variables. The two clustering methods we used were the k-means method and the hierarchical clustering method. For each clustering method, we had two different analyses that focused on student characteristics and instructor characteristics. To interpret our categories for the clustering analyses, we calculated the means of each section for each variable and then standardized each variable. We also only included sections where the concept test was administered on both occasions of pre and post. After conducting cluster analyses with both clustering methods, we decided to use clustering variables based on the k-means method, as the number of sections for each cluster was more equal compared to the hierarchical clustering method.

Instructor classification: Ad-hoc classification

For the ad-hoc method, we established four different categories that were experienced (e.g., members of the author team), middle, new (this was their first time using the materials), and non-users (we had two sections where the instructor was not using the ISI materials but agreed to administer the same assessments). Below are two graphs showing the relationship between improvement and the level of experience with the curriculum. The left graph shows the variability in improvements among the four categories while the right graph shows the variability within each of the four categories, equivalently among sections of the same category.



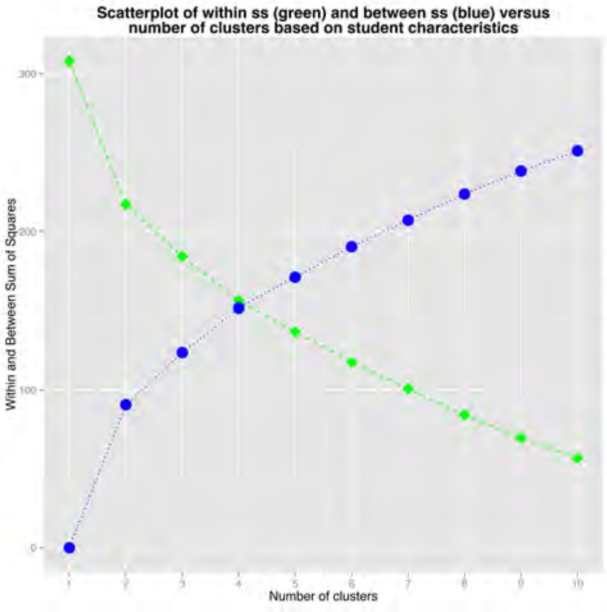


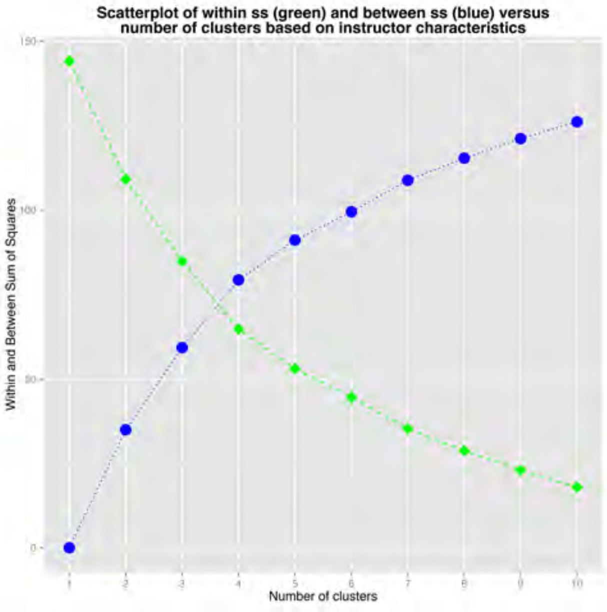
From the graph on the left, non-users tended to have students who did not improve as much as the three other categories of users, on average. Also, students who were taught by instructors more experienced with the curriculum tended to have the best improvements. The graph on the right illustrates how sections differ in each category. It is shown that there are considerable differences among sections, especially for the experienced users and new users; sections did not differ as much on average for middle users and non-users as the variability and means of the sections did not generally differ that much. With the information portrayed from these two graphs, we hypothesized that there may be a significant difference between experienced users and non-users, while users of the other levels may not be significantly different in improvements.

Instructor classification: K-means

Most "optimal" number of clusters

To decide on the most "optimal" number of clusters to use, we studied how the within and between sum of squares of the student and instructor level variables changed with different number of clusters based on the k-means method. Below are two graphs that show the within sum of squares in green and the between sum of squares in blue. Student and instructor characteristics are shown on the left and right graphs, respectively.

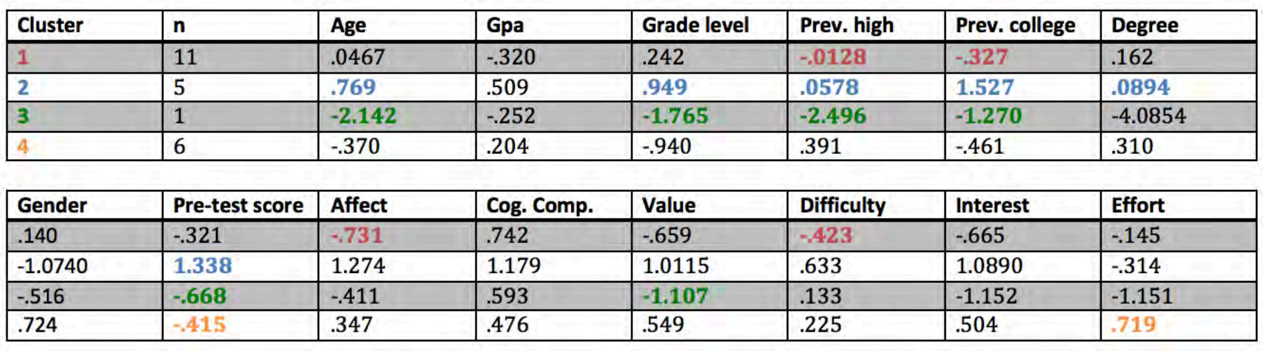




The results of the above two graphs made us understand that the between sum of squares (in blue) started to exceed the within sum of squares (in green) when the number of clusters was roughly around four. In a mathematical sense, it may be interpreted that the most optimal number of clusters to use would be ten because the largest difference between the two sums of squares occurred at the number. However, our decision was based on a statistical perspective where we did not want multiple clusters to only consist of one instructor. It is also worth noting that with four clusters, we already had a cluster with only one instructor in it. Therefore, we chose to use four clusters as the number was a "balancing point" of the two types of sum of squares.

Clustering based on student characteristics

We then ran our first k-means cluster analysis based on 14 student level variables, which described student characteristics. Such student characteristics were their age, GPA, grade level, number of previous high school and college classes related to mathematics and statistics, degree seeking (higher positive numbers indicated higher degree seeking), gender (higher positive numbers indicated more females), pre-test scores, and 6 attitudes variables (as measured by taking the average of corresponding questions from the attitudes test). Below is the table of the means of each variable for each cluster. Certain means were color coded when those clusters stood out.

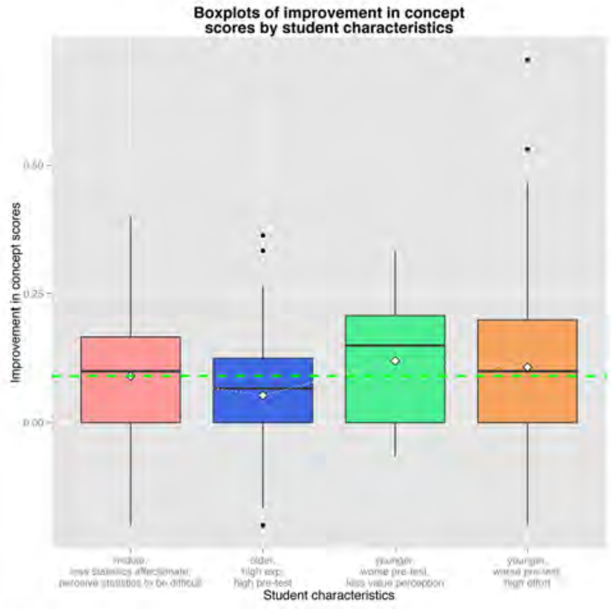


Cluster	n	Age	Gpa	Grade level	Prev. high	Prev. college	Degree	Gender	Pre-test score	Affect	Cog. Comp.	Value	Difficulty	Interest	Effort
1	11	.0467	-.320	.242	-.0128	-.327	.162	.140	-.321	-.731	.742	-.659	-.423	-.665	-.145
2	5	.769	.509	.949	.0578	1.527	.0894	-1.0740	1.338	1.274	1.179	1.0115	.633	1.0890	-.314
3	1	-2.142	-.252	-1.765	-2.496	-1.270	-4.0854	-.516	-.668	-.411	.593	-1.107	.133	-1.152	-1.151
4	6	-.370	.204	-.940	.391	-.461	.310	.724	-.415	.347	.476	.549	.225	.504	.719

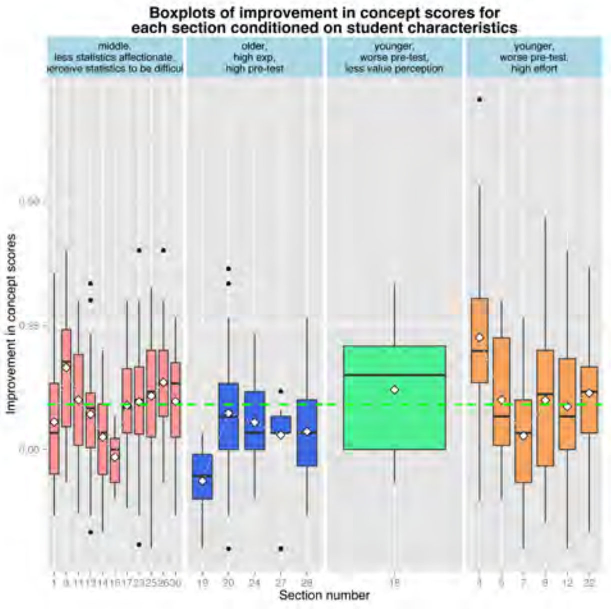
We proceeded to characterize each cluster with the main features observed in the table. Our classification of the four clusters was as follow:

- Cluster 1: Instructors with students who tended to have more negative feelings about statistics and its difficulty and had less prior experience with statistics in high school or college going into the course.
- Cluster 2: Instructors with older students who tended to have more previous exposure to statistics and scored high on the pre-test. These students also tended to have better pre-attitudes about statistics.
- Cluster 3: Instructors with younger students (high-school) who tended to have less prior experience in statistics and lower perception of value of statistics. The students also tended to score the lowest on the pre-test, did not perceive to exert much effort, and had lower interests in statistics.
- Cluster 4: Instructors with younger students who tend to have students who did worse on the pre-test but had high expected effort.

After characterizing the four clusters, we generated boxplots to show how the four clusters differed in respect to improvement in concepts along with how sections in each cluster differed. The means of each boxplot are shown in diamonds and the boxplots were color coded according to each of the four clusters.



Boxplots of improvement in concept scores by student characteristics



Boxplots of improvement in concept scores for each section conditioned on student characteristics

From the graph on the left, it was intriguing that the instructor with younger (high-school) students who had worse pre-test scores and lower perception of value of statistics (green boxplot) tended to have higher improvements than the other three clusters. We also found that instructors with younger students who were weaker but had high expected effort (brown boxplot) also tended to have higher improvements. This relationship may be due to their ethic to work harder as they may have realized that they did not understand statistics as much after taking the pre-test. In contrast, instructors with older students who scored higher on the pre-test (blue boxplot) tended to not improve as much. One possible explanation is that they may have already done so well on the pre-test that after taking the course their knowledge of statistics did not broaden as much compared to students with less knowledge going into the course. However, the boxplots of each cluster still have some considerable overlapping depicting not as much between-cluster variability.

When analyzing the within-cluster variability based on the graph on the left, there is considerable variability among sections in each cluster. For example, section 3 in the "middle" category has a higher average than the other sections in the same category, section 19 in the "older" category has a lower average than the other sections in the same category, and section 4 in the "younger with worse pre-test" category has a higher average than the other sections in the same category.

Clustering based on instructor characteristics

Our second k-means cluster analysis was based on 8 instructor level variables, which described instructor, section, and school characteristics. Such variables were department (higher positive numbers indicated closer to being in the statistics department), tenure status (higher positive numbers indicated closer to being tenured), school type was coded (higher positive numbers indicated the school was closer to a four-year university, years of teaching, student and instructor-led involvement, length of weeks, and time of day (higher positive numbers indicated the class being later in the day). Below is the table of the means of each variable for each cluster. Certain means were color coded when those clusters stood out.

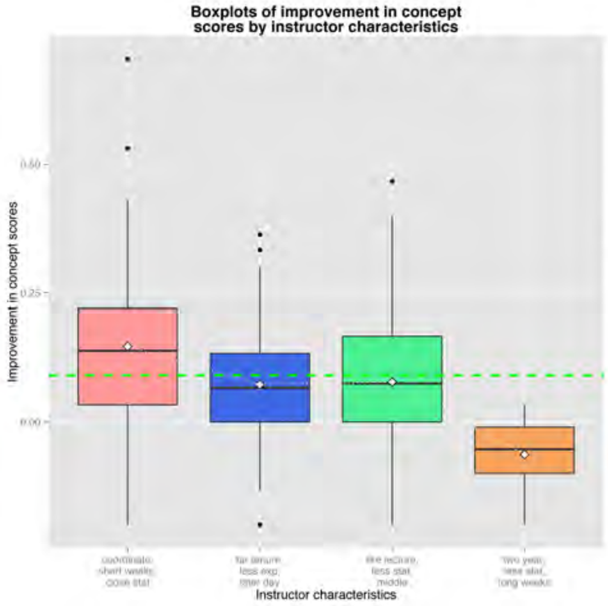
Cluster	n	Department	Tenure status	School type
1	4	1.197	-.265	.229
2	6	-.652	-.602	.229
3	8	-.0511	.491	.229
4	1	-.467	.743	-4.129

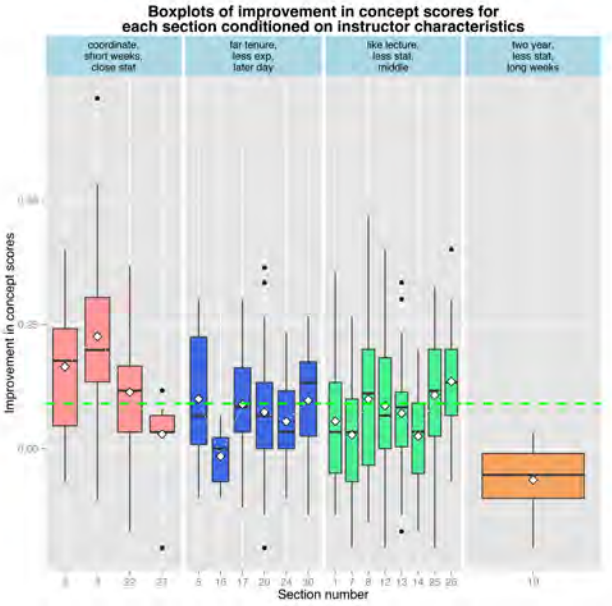
Years teaching	Student led	Instructor led	Length of weeks	Time of day
.204	1.246	-.992	-1.331	-.491
-.432	.428	-.319	.126	.972
.134	-.811	.927	.467	-.354
.709	-1.0633	-1.531	.834	-1.0396

We proceeded to characterize each cluster with the main features observed in the table. Our classification of the four clusters were as follow:

- Cluster 1: Instructors who incorporated more student-led involvements, tended to be in the statistics department, and were on quarter system.
- Cluster 2: Instructors who were newer with less experience in teaching and had classes that met later in the day.
- Cluster 3: Instructors who liked to lecture, tended to not be in the statistics department.
- Cluster 4: A two-year college instructor from the math department who had a class that met for longer weeks and earlier in the day.

To illustrate how the four clusters differed in improvement in concepts along with the disparity in the sections of each cluster, boxplots were created. The means of each boxplot are plotted in diamonds and the boxplots were color coded according to each of the four clusters.





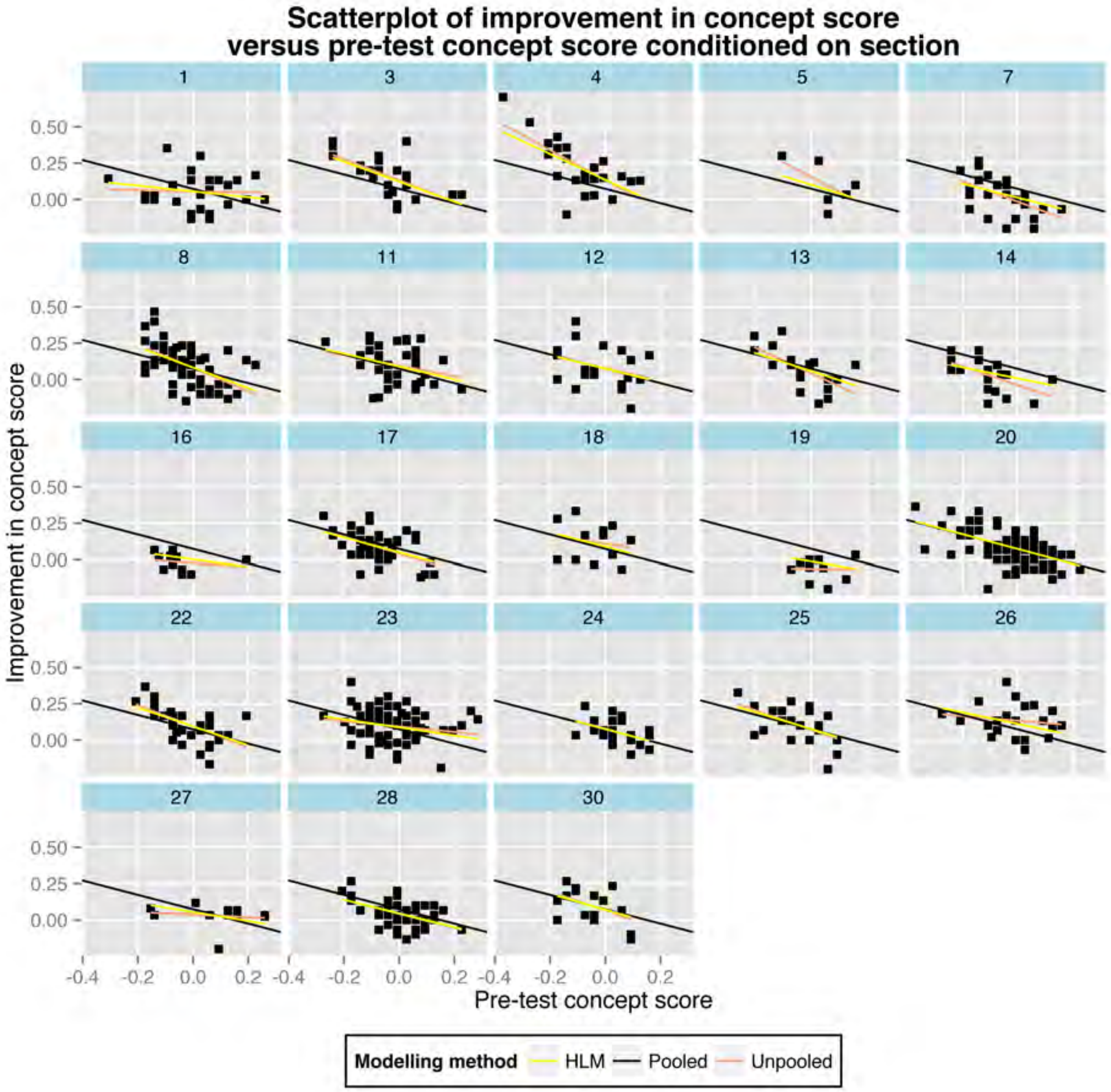
The graph on the left shows that instructors who had more student involvements (red boxplot) tended to have a higher improvement in concepts compared to instructors who liked to lecture (green boxplot). Also, instructors with less experience in teaching (blue boxplot) tended to have lower improvement in concepts except when compared to the two-year college instructor. The two-year college instructor from the math department (brown boxplot) had the worst improvement in concepts on average than the other three clusters. There is more variability in improvements among the four clusters based on instructor characteristics as compared to earlier based on level of experience with curriculum and student characteristics.

The graph on the right depicts the how sections within each cluster varied. It is evident that all except for section 27 in the “more student-involvement“ cluster were above the national average. The section in the “less experience in teaching” cluster that is pulling the average of the cluster down is section 16. In the “like lecture” cluster, the sections were generally the same in terms of the variability but slightly different means. Lastly, the two-year instructor had an average improvement that is lower than the national’s and is below zero.

Comparison of Modelling Methods

Partial pooling and borrowing strength

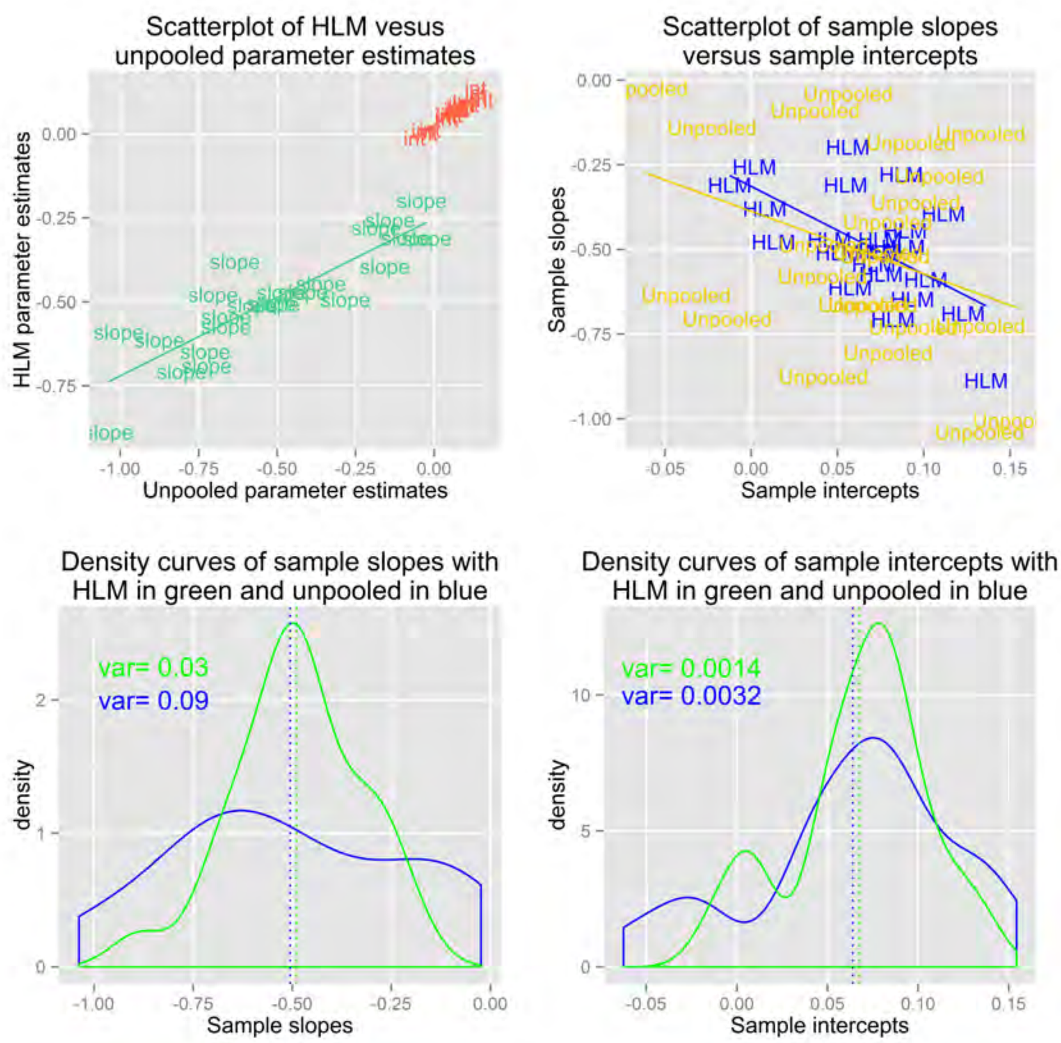
We focused on comparing the fitted lines among the three modelling methods. For the HLM, we have fitted lines for each instructor after predicting improvement in concepts with pre-test concept as the random predictor. The complete pooling method involved regressing improvement in concepts on pre-test concept while ignoring instructor-to-instructor variability. The no pooling method fitted the previously mentioned regression but for each individual instructor. With the fitted lines, we studied how the HLM lines differed among the sections that we had. The graph on the next page shows a scatterplot of improvement in concepts versus pre-test concept separated by each section. The complete pooling line is plotted in black and did not change across sections, as we ignored the instructor-to-instructor variability. The no pooling lines are plotted in red and the HLM lines are plotted in yellow, and these fitted lines did change across sections.



We learned from the graph above the HLM served as a balance between the two extreme methods of complete pooling and no pooling. It is evident that for each section, the HLM line was always between the pooled line and unpooled line. Therefore, HLM is a partial pooling technique where each section borrows information from the overall trend. The only difference across the sections is the amount of borrowing strength. As expected, we see that borrowing strength increased when there is a smaller sample size in a section and a section with a larger sample size did not have to borrow as much information from the other sections. For example, comparing section 4 and 5 illustrated the above ideas. The yellow HLM line for section 5 is closer to the black pooled line than to the red unpooled line as there were fewer students in that section. In contrast, the yellow HLM line for section 4 is closer to the red unpooled line than to the black pooled line as there were more students in that section.

Shrinkage

Next, we compared the variability of the unpooled estimates to the HLM estimates. Scatterplots of the parameter estimates and the distribution of the parameter estimates were created to compare the two different modelling methods.



The upper-left graph shows a scatterplot of the HLM parameter estimates versus the unpooled parameter estimates with the sample slopes in green and sample intercepts in red. It is evident that the unpooled estimates are more variable than the HLM estimates, as the range of x-values is larger than the range of y-values for both the sample slopes and sample intercepts. The upper-right graph shows a scatterplot of the sample slopes versus sample intercepts identified either estimated through HLM or the unpooled method. Evidently, the HLM estimates tend to be more cluttered together compared the unpooled estimates. Also, as the sample intercepts increased, the sample slopes tended to decrease. This relationship made sense because a more negative association between improvement in concepts and pre-test concept for each section is related higher predicted improvement in concepts when the pre-test concept was zero. The two graphs below show the distributions of the sample slopes and sample intercepts in green and blue for the HLM and unpooled method, respectively. It is evident that the idea of “shrinkage” in the parameter estimates is shown, as the variability in the sample slopes is about

3 times lower with the HLM compared to the unpooled method. Also, the variability in the sample intercepts is about a little above 2 times higher with the HLM compared to the unpooled method.

Confirmatory Data Analysis

Unconditional means model with improvement in concepts as response

Proceeding to running our hierarchical models, we first fitted an unconditional means model to quantify the differences among instructors in their improvement in concepts. We calculated the intraclass correlation coefficient/variance partition coefficient with the formula:

$$\hat{\rho} = \frac{Between - section\ variability\ in\ improvement\ in\ concepts}{Total\ variability\ in\ improvement\ in\ concepts}$$
$$= \frac{\hat{\sigma}_0^2}{\hat{\sigma}_0^2 + \hat{\sigma}^2} = \frac{.002568}{.002568 + .0140} = .154$$

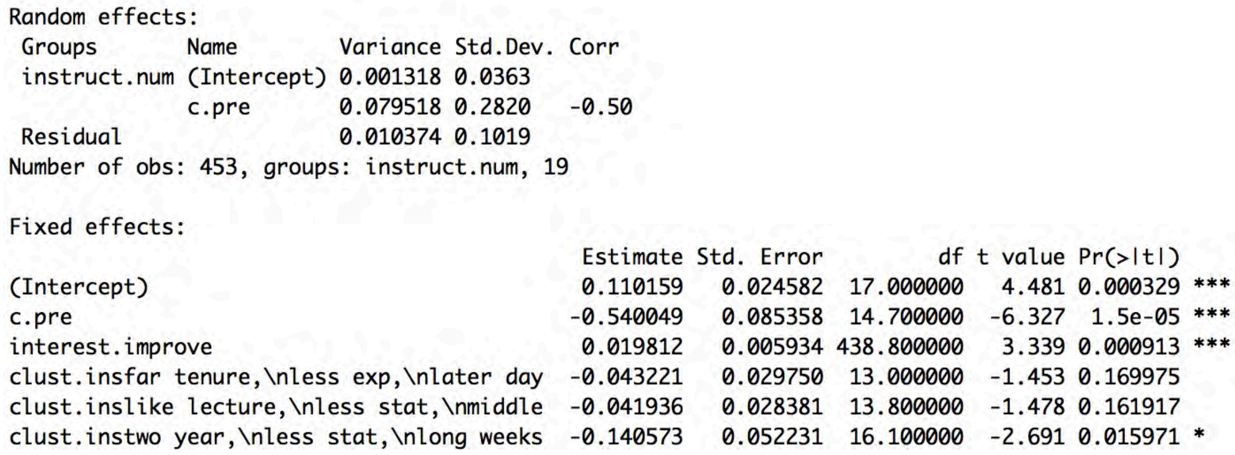
This estimate can be interpreted in two different ways as it is considered to represent two ideas:

1. The heterogeneity among instructors/sections is estimated to account for 15.4% of the total variability in improvement in concepts
2. A pair of two students within the same instructor/section is estimated to have a correlation of .154

Despite either interpretation, we have found that there is not much variability among instructors as the variability within instructor is much larger. Past education research have reported an intraclass correlation coefficient for similar analysis to be around 10%. Therefore, we decided to continue our analysis by defining the hierarchy in the model in respect to the structure of the data. We explored numerous models after the unconditional means model but only the most interesting ones is discussed in the following sections.

Model 1

The first model that we will discuss in detail was a model predicting improvement in concepts with the predictors of the clustering based on instructor characteristics and student level variables. In total, there were 453 students and 19 sections that did not have missing data for the variables involved. Shown is the output for the model with the 2 significant student-level predictors of concepts pre-test and improvement in interest toward statistics and the 1 significant instructor-level predictor of clustering based on instructor characteristics. We ended up with this model after first fitting a model with many predictors and then paring it down with backward elimination after all predictors were found to be significant. The AIC value also resulted to be -715.



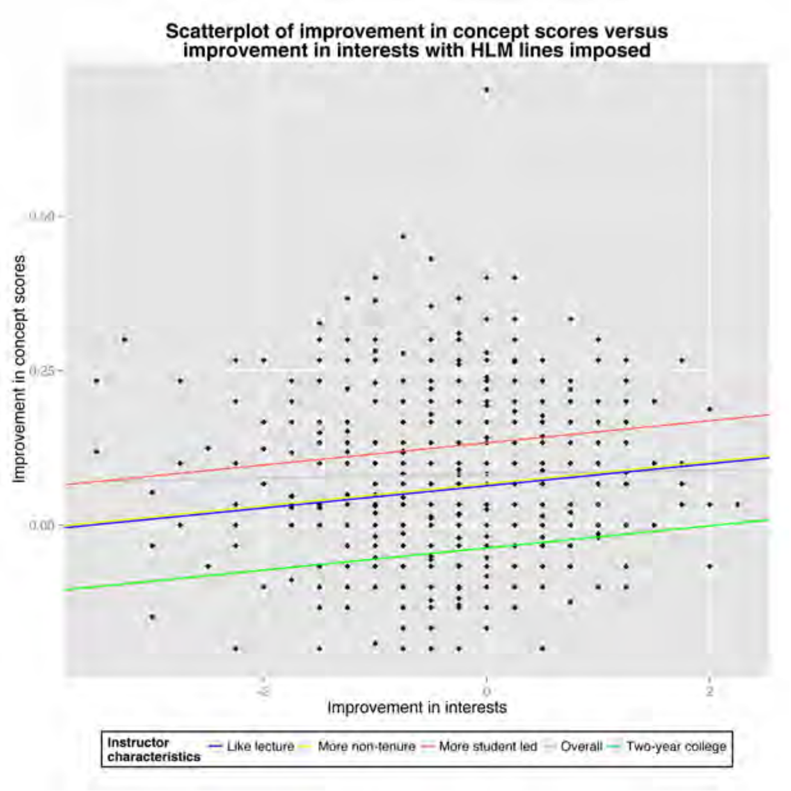
```
Random effects:
 Groups       Name        Variance Std.Dev. Corr
 instruct.num (Intercept) 0.001318 0.0363
              c.pre       0.079518 0.2820   -0.50
 Residual                 0.010374 0.1019
Number of obs: 453, groups: instruct.num, 19

Fixed effects:
                                             Estimate Std. Error         df t value Pr(>|t|)
(Intercept)                                  0.110159   0.024582  17.000000   4.481 0.000329 ***
c.pre                                       -0.540049   0.085358  14.700000  -6.327  1.5e-05 ***
interest.improve                             0.019812   0.005934 438.800000   3.339 0.000913 ***
clust.insfar tenure,\nless exp,\nlater day  -0.043221   0.029750  13.000000  -1.453 0.169975
clust.inslike lecture,\nless stat,\nmiddle  -0.041936   0.028381  13.800000  -1.478 0.161917
clust.instwo year,\nless stat,\nlong weeks  -0.140573   0.052231  16.100000  -2.691 0.015971 *
```

It is evident that after adjusting for improvement in interest and instructor characteristics, higher pre-test concepts are associated with a lower improvement in concepts. This association reflects that students who did worse on the pre-test tended to improve more while students who did better on the pre-test did not have as much room for improvements. The only significant student attitudes predictor with this model was improvement in interests. The association between improvement in interests and improvement in concepts is positive, after adjusting for the other two variables in the model. Therefore, increasing students' interest in statistics from pre-test to post-test is associated with them having higher improvements in concepts, after adjusting for their pre-test concepts and the type of instructor they had. In regards to instructor characteristics, the only statistically significant difference is between the two-year college instructor and the "more student-led involvement" category of instructors. The coefficient also told us that being taught by the two-year instructor is associated with an average drop of 15 percentage points from the concepts pre-test to the post-test compared to being taught by instructors in the "more student-led involvement" category, after adjusting for concepts pre-test and how much interest in statistics has been improved.

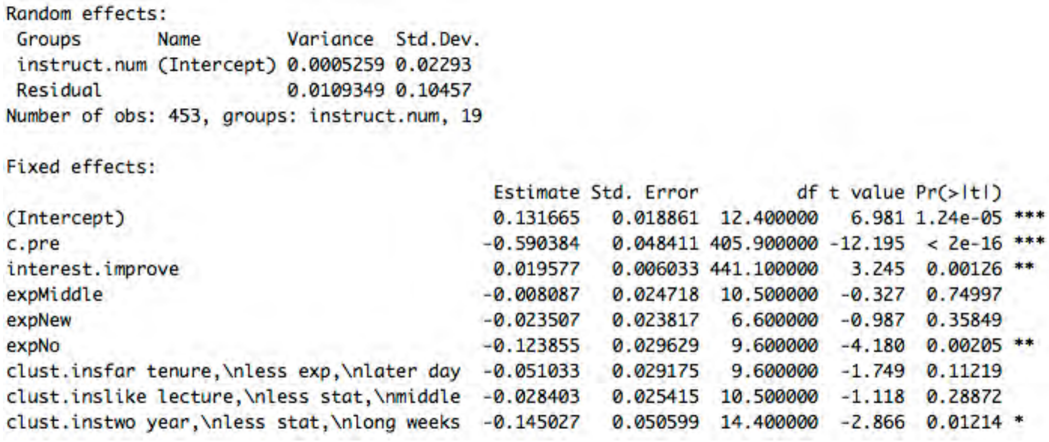
To summarize this model in a practical sense, we see a student will have large improvements in the concept test if they score lower on the pre-test, is taught by an instructor with more student-level involvement, and has an improvement in interest in statistics in the course.

We also generated a scatterplot of improvement in concepts versus improvement in interests with the HLM fitted lines for each category of instructors imposed. (Note: the interaction was not statistically significant.) The scatterplot shows that the red fitted line for the "more-student led" cluster is higher than all the other three clusters. The two clusters of "more non-tenure" and "like lecture" are very similar to each other, lower than "more-student led", and higher than "two-year college." The green fitted line for the "two-year college" instructor is the lowest of all cluster fitted lines. However, the association of improvement in concepts and improvement in interests is the same across all clusters.



Adding the variable of experience with curriculum to Model 1

We were also interested in determining whether there would be a significant difference among instructors of different level of experience with the curriculum, after adjusting for the significant predictors from model 1. Therefore, we added the ad-hoc experience variable into model 1. The AIC of the model resulted in -702, meaning this model did not perform quite as well as the previous model. We found that non-users tend to have a significantly lower improvement in concepts than experienced users of the curriculum, after adjusting for other instructor characteristics and student improvement in interests. The output of this model is shown below.



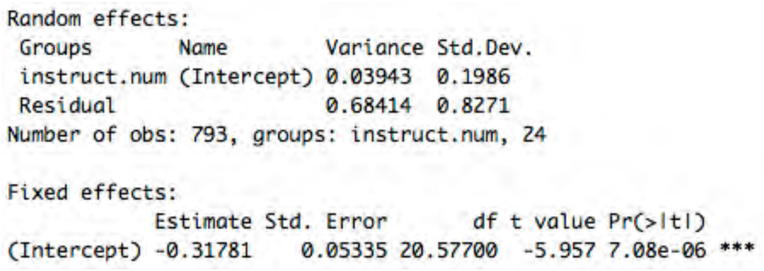
```
Random effects:
 Groups       Name        Variance  Std.Dev.
 instruct.num (Intercept) 0.0005259 0.02293
 Residual                 0.0109349 0.10457
Number of obs: 453, groups: instruct.num, 19

Fixed effects:
                                           Estimate Std. Error        df t value Pr(>|t|)
(Intercept)                                0.131665   0.018861  12.400000   6.981 1.24e-05 ***
c.pre                                     -0.590384   0.048411 405.900000 -12.195  < 2e-16 ***
interest.improve                           0.019577   0.006033 441.100000   3.245  0.00126 **
expMiddle                                 -0.008087   0.024718  10.500000  -0.327  0.74997
expNew                                    -0.023507   0.023817   6.600000  -0.987  0.35849
expNo                                     -0.123855   0.029629   9.600000  -4.180  0.00205 **
clust.insfar tenure,\nless exp,\nlater day -0.051033   0.029175   9.600000  -1.749  0.11219
clust.inslike lecture,\nless stat,\nmiddle -0.028403   0.025415  10.500000  -1.118  0.28872
clust.instwo year,\nless stat,\nlong weeks -0.145027   0.050599  14.400000  -2.866  0.01214 *
```

Predicting improvement in interests

As improvement in interests was the significant student-level attitudes predictor, we wanted to determine the significant predictors of this variable. The purpose was to see if we could identify which types of students or instructors are associated with higher improvement in interests, then they would also be linked with higher improvement in concepts.

We first ran an unconditional means model to obtain the ICC/VPC. Below is the output of the unconditional means model.

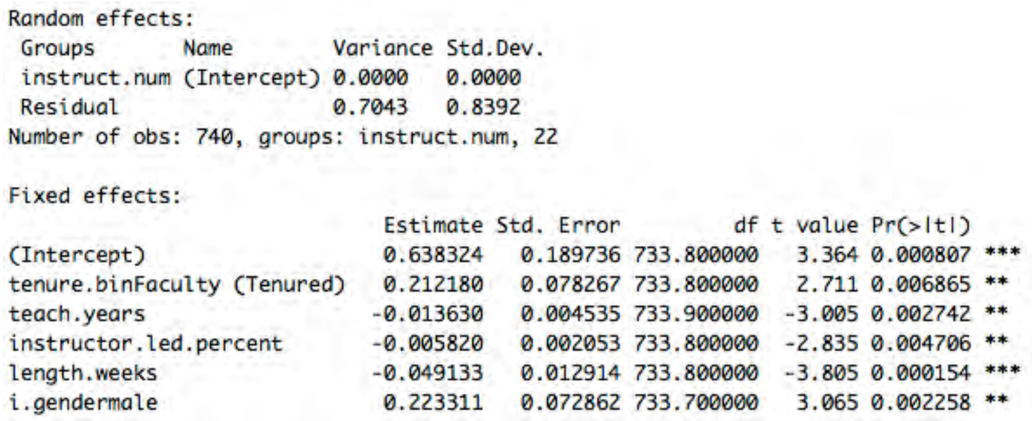


```
Random effects:
 Groups       Name        Variance Std.Dev.
 instruct.num (Intercept) 0.03943  0.1986
 Residual                 0.68414  0.8271
Number of obs: 793, groups: instruct.num, 24

Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)
(Intercept) -0.31781    0.05335 20.57700  -5.957 7.08e-06 ***
```

The estimated intraclass correlation coefficient turned out to be .0545 after calculations were done. This small value revealed that differences among instructors/sections did not account much for the total variability in improvement in interests.

We started with a model including all instructor-level predictors and student-level predictors that were not based on the attitudes pre-test. The model that we finalized on after paring was done included the 5 significant instructor-level predictors of whether or not tenured, years of teaching, instructor-led percent, length of classes, and gender of instructor. Below is the output of the model fitted and the AIC was 1874.



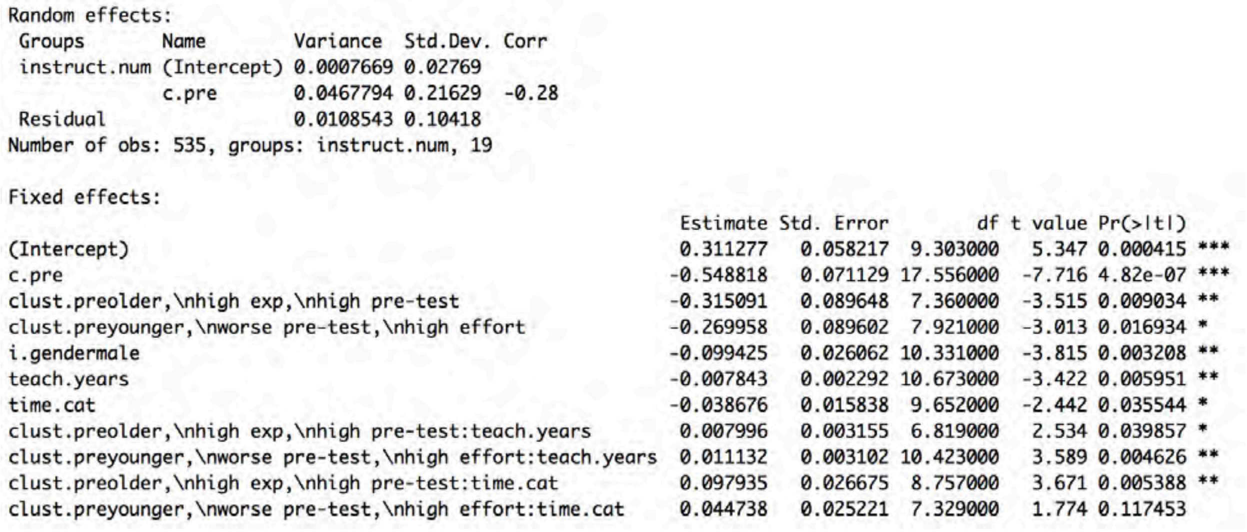
```
Random effects:
 Groups       Name        Variance Std.Dev.
 instruct.num (Intercept) 0.0000   0.0000
 Residual                 0.7043   0.8392
Number of obs: 740, groups: instruct.num, 22

Fixed effects:
                            Estimate Std. Error         df t value Pr(>|t|)
(Intercept)                 0.638324   0.189736 733.800000   3.364 0.000807 ***
tenure.binFaculty (Tenured)  0.212180   0.078267 733.800000   2.711 0.006865 **
teach.years                -0.013630   0.004535 733.900000  -3.005 0.002742 **
instructor.led.percent     -0.005820   0.002053 733.800000  -2.835 0.004706 **
length.weeks               -0.049133   0.012914 733.800000  -3.805 0.000154 ***
i.gendermale                0.223311   0.072862 733.700000   3.065 0.002258 **
```

From this model, the most ideal instructor, for student improvement in interest in statistics, is a tenured male instructor who has moderate experience of teaching and focuses on student-led involvement in a school on quarter system.

Model 2

The second model that we fitted was a model predicting improvement in concepts with the predictors of the clustering based on student characteristics and instructor-level variables. In total, there were 535 students and 19 sections that did not have missing data for the variables involved. Shown is the output for the model with the 1 significant predictor of concepts pre-test and the 6 significant instructor-level predictors of the clustering variable based on student characteristics, gender of instructor, years of teaching, time of the day, the interaction of the clustering variable and years of teaching, and the interaction of the clustering variable and time of the day. The cluster of instructors with younger students who had worse pre-test and less value perception was dropped from the model as we had missing data on the cluster for at least one of the predictors in the model. We ended up with this model after first fitting a model with many predictors and then paring it down with backward elimination after all predictors were found to be significant. The AIC value also resulted to be -799.



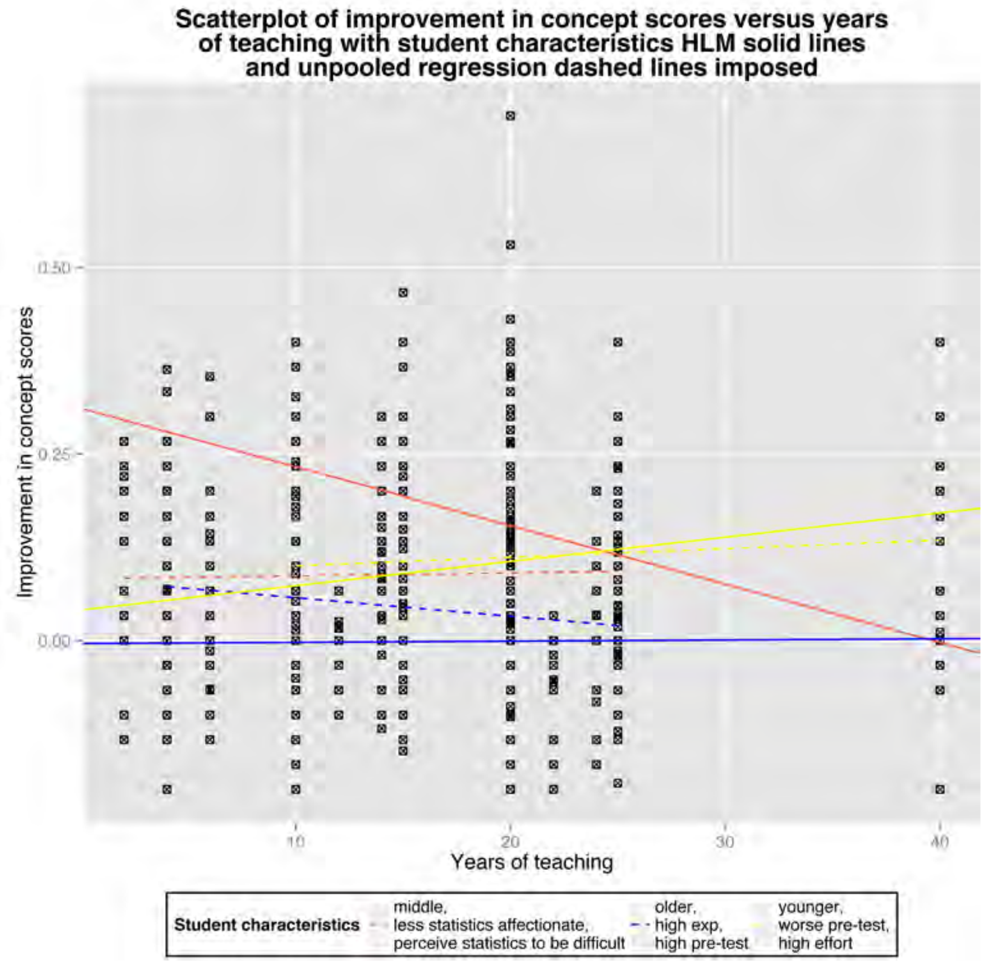
```
Random effects:
 Groups       Name        Variance  Std.Dev. Corr
 instruct.num (Intercept) 0.0007669 0.02769
              c.pre       0.0467794 0.21629  -0.28
 Residual                 0.0108543 0.10418
Number of obs: 535, groups: instruct.num, 19

Fixed effects:
                                                         Estimate Std. Error        df t value Pr(>|t|)
(Intercept)                                              0.311277   0.058217  9.303000   5.347 0.000415 ***
c.pre                                                   -0.548818   0.071129 17.556000  -7.716 4.82e-07 ***
clust.preolder,\nhigh exp,\nhigh pre-test               -0.315091   0.089648  7.360000  -3.515 0.009034 **
clust.preyounger,\nworse pre-test,\nhigh effort         -0.269958   0.089602  7.921000  -3.013 0.016934 *
i.gendermale                                            -0.099425   0.026062 10.331000  -3.815 0.003208 **
teach.years                                             -0.007843   0.002292 10.673000  -3.422 0.005951 **
time.cat                                                -0.038676   0.015838  9.652000  -2.442 0.035544 *
clust.preolder,\nhigh exp,\nhigh pre-test:teach.years    0.007996   0.003155  6.819000   2.534 0.039857 *
clust.preyounger,\nworse pre-test,\nhigh effort:teach.years 0.011132 0.003102 10.423000 3.589 0.004626 **
clust.preolder,\nhigh exp,\nhigh pre-test:time.cat       0.097935   0.026675  8.757000   3.671 0.005388 **
clust.preyounger,\nworse pre-test,\nhigh effort:time.cat 0.044738   0.025221  7.329000   1.774 0.117453
```

The most intriguing findings from this model were from the interaction terms. For the interaction of the clustering variable and years of teaching, we found that for students who had more prior experience in statistics and did better on the pre-test improvement in concepts did not tend to be associated with the years of teaching of the instructor, after adjusting for other predictors in the model. For students who performed worse on the pre-test but had high expected effort, their improvement in concepts tend to increase with more years in teaching, after adjusting for other predictors in the model. Also, students with more negative tended to decline with more experienced instructors, after adjusting for other predictors. The second interaction of the clustering variable and time of day provided evidence that if the class met later in the day, the improvement in concepts tend to increase for both clusters of instructors who had high experience students and students who did worse on the pre-test but had high expected effort. In contrast, for the cluster of instructors who had the “middle” students, if the cluster met in the day, the improvement in concepts tended to decrease.

For the two remaining significant main effects, after adjusting for the predictors in the model, students with male instructors tended to do worse than students with female instructors. Also, after making the necessary adjustments, students who did well on the concepts pre-test tended to have lower improvements.

The ideal student with this model is a high effort student taught by an experienced instructor in a course meeting later in the day.



Above is a scatterplot with the HLM and unpooled lines of the clustering of instructors based on student characteristics imposed to illustrate one of the ideas from the model fitted. It is evident that there is some disparity between the HLM and unpooled lines so accounting for the hierarchy does make a difference in the associations. The yellow line is for instructors who had younger students with worse pre-test but high effort. There is a positive slope for this line, which agrees with the earlier interpretation that students with lower pre-test scores tend to show higher improvements. Also instructors with students who were older and high previous experience in statistics did not have an association between improvement in concepts and years of teaching, as the blue line is pretty flat. For the red line, the “middle” students who had instructors with more years of teaching tended to have lower improvement in concepts than those with newer teachers.

Conclusion

Limitations of research

We hope to gather more data from Spring 2014 in the future for many reasons that restrained our data analysis. One of the reasons was that we only had 2 non-users compared to the 28 users of the curriculum so recruiting more non-users can increase our sample size for that category. We also decided to not include a three-level hierarchy in running our hierarchical models due to the lack of information on the institutions that the instructors were from and the limited number of instructors for some institutions. One approach is to widen our scope by contacting more instructors who teach statistics at an introductory level. After having more data, we are interested to defining the third level of the hierarchy to not only identify the student and instructor characteristics associated with student gains but also the institution characteristics. We are aware that this research is only a start to a much more extensive research to incorporate more information from institutions and instructors across the nation.

In addition, we currently have data on students from winter 2014 who took the assessment tests under 10 minutes. Therefore, we wish to exclude these students from the analysis as these type of students were already excluded in the fall 2013 data.

Goals achieved

Our two primary goals from the beginning of this research were to identify the student and instructor characteristics associated with students gains of the randomization-based curriculum and to detect whether there was a disparity in improvement in concepts among instructors who were at different level of experience with the curriculum. We achieved our first primary goal by running multiple models. Based on the model after adding the experience with curriculum variable to Model 1 and the model to predict improvement in interest, we found that the most "ideal" student who would on average have the highest gains is a student who scores low on the concepts pre-test, increases their interest from pre-test to post-test, and is taught by a moderate teaching experience male instructor who has high experience with the curriculum and includes more student-led involvements. However we did find some different results after running Model 2 where the most "ideal" students is a high effort student taught by an experience instructor who met later in the day. The second primary goal was also achieved after we added the experience with curriculum level into Model 1. We found that after adjusting for other instructor characteristics, students' concepts pre-test performance, and students' improvement in interest, experienced users tended to have significantly higher improvements than non-users. In addition, there was no evidence that users of the curriculum at different levels were significantly different. The two secondary goals that were to understand and apply hierarchical models and to utilize *R* for performing analyses and producing graphical displays were also reached as we found the HLM to be a partial pooling method involving the ideas of borrowing strength and shrinkage and we conducted different analyses and produced many graphs in *R* throughout our research.

References

Chance B. & Rossman A. (2014). *Investigating Statistical Concepts, Applications, and Methods.*

Garfield J. (1995). *How Students Learn Statistics.*

Garfield J. & Ben-Zvi D. (2009). *How Students Learn Statistics Revisited: A Current Review of Research on Teaching and Learning Statistics.*

Lock RH, Lock PF, Lock Morgan K, Lock EF, Lock DF. *Statistics: Unlocking the Power of Data*. Hoboken, NJ: John Wiley and Sons; 2013.

Rossman A. & Chance B. (2008). www.rossmanchance.com/applets

Tintle N., VanderStoep J., Holmes V., Quisenberry B. & Swanson T. (2011). *Development and assessment of a preliminary randomization-based introductory statistics curriculum.*

Tintle N., Topliff K., Vanderstoep J., Holmes V. & Swanson T. (2012). *Retention of Statistical Concepts in a Preliminary Randomization-based Introductory Statistics Curriculum.*